IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 February 2024, accepted 21 March 2024, date of publication 27 March 2024, date of current version 15 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381967

SURVEY

A Survey on Goal-Oriented Semantic Communication: Techniques, Challenges, and Future Directions

TILAHUN M. GETU[1], **(Member, IEEE), GEORGES KADDOUM**[1,2], **(Senior Member, IEEE), AND MEHDI BENNIS**[3], **(Fellow, IEEE)**

[1]Electrical Engineering Department, ÉTS, Montreal, QC H3C 1K3, Canada
[2]Cyber Security Systems and Applied AI Research Center, Lebanese American University, Beirut 1102-2801, Lebanon
[3]Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland

Corresponding author: Tilahun M. Getu (tilahun-melkamu.getu.1@ens.etsmtl.ca)

This work was supported by the U.S. Department of Commerce and its agency National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. The work of Tilahun M. Getu was supported by NIST [Dr. Hamid Gharavi (Life Fellow, IEEE)].

ABSTRACT Although many proposals have been developed for the sixth-generation (6G) technology, realizing 6G is fraught with numerous fundamental interdisciplinary, multidisciplinary, and transdisciplinary challenges. To mitigate some of these challenges, goal-oriented semantic communication (SemCom) has emerged as a promising 6G technology enabler. This enabler employs only semantically-relevant information for successful task execution while minimizing power usage, bandwidth consumption, and transmission delay. On the other hand, 6G is essential for realizing major goal-oriented SemCom use cases such as autonomous transportation. These paradigms of *6G for goal-oriented SemCom* and *goal-oriented SemCom for 6G* call for a tighter integration of 6G and goal-oriented SemCom. To facilitate this purpose, this survey paper exposes the fundamental challenges of 6G; details the notion of goal-oriented SemCom and its state-of-the-art research landscape; presents state-of-the-art trends, use cases, and frameworks of goal-oriented SemCom; exposes the fundamental and major challenges of goal-oriented SemCom; and offers promising future research directions for goal-oriented SemCom. Consequently, this survey article stimulates numerous lines of research on goal-oriented SemCom theories, algorithms, and realization.

INDEX TERMS 6G, goal-oriented SemCom, techniques of goal-oriented SemCom, challenges of goal-oriented SemCom, future directions for goal-oriented SemCom.

I. INTRODUCTION

A. CONTEXT

In light of the contemporary fifth-generation (5G) technology, it is widely anticipated that 5G new radio (5GNR) [1], [2] and its subsequent releases will offer enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine type communications (mMTC) services to the 2020s society. Nonetheless, to support the proliferation of *internet of everything (IoE)* services, such as extended reality (XR) [3], a wireless communication network should be able to simultaneously support ultra-high data rates, ultra-low transmission latency, and hyper-connectivity [4]. In this vein, it is debatable whether 5G and its future evolution could support the simultaneous requirements of the eMBB, URLLC, and mMTC services, which is arguably 5G's fundamental limitation that should be overcome in the sixth-generation (6G) technology.

Amidst the global rollout of 5G, researchers in academia, industry, and national laboratories have been developing vision for 6G wireless communication system technology [3], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16],

The associate editor coordinating the review of this manuscript and approving it for publication was Nafees Mansoor.

TABLE 1. Driving 6G applications along with their potential for industry verticals of e-health, automotive, energy, media and entertainment, factory of the future, and/or transportation: ✓and × symbolize yes and no, respectively, to a given industry vertical.

	Industry Verticals					
6G Driving Applications	**e-Health**	**Automotive**	**Energy**	**Media and Entertainment**	**Factory of the Future**	**Transportation**
Haptic communication [35]	✓	×	×	✓	×	×
Massive IoT integrated smart city [35]	✓	✓	✓	✓	✓	✓
Automation and manufacturing [35]	×	×	✓	×	✓	✓
Multisensory extended reality (XR) applications [3]	✓	×	×	✓	×	×
Connected robotics and autonomous systems [3]	✓	✓	×	×	×	✓
Brain-computer interaction/interface (BCI) [3]	✓	×	×	✓	×	×
Blockchain and distributed Ledger technologies [3]	✓	×	×	×	✓	✓
The internet of no things (the Metaverse) [31], [32]	✓	✓	×	✓	✓	✓
Industrial IoT [20], [36]	×	×	×	×	✓	×
Internet of robots [28]	×	✓	×	×	✓	✓
Flying vehicles [21]	×	✓	×	×	×	✓
Wireless data centers [21], [37]	✓	✓	✓	✓	✓	✓
Accurate indoor positioning [15]	✓	✓	×	✓	×	✓
New communication terminals [15]	✓	×	×	✓	×	×
High-quality communication services on board aircraft [15]	✓	×	×	✓	×	×
Worldwide connectivity and integrated networking [15]	✓	×	×	✓	×	×
Communications that support industry verticals [15]	✓	✓	✓	✓	✓	✓
Holographic and tactile communications [15]	✓	✓	×	✓	✓	✓
Human-bond communications [15]	✓	×	×	×	×	×
Smart Grid 2.0 [6]	×	×	✓	×	✓	×
Industry 5.0 [6]	×	×	×	×	✓	×
Personalized body area networks [6]	✓	×	×	×	×	×
Healthcare 5.0 (digital wellness) [6]	✓	×	×	×	×	×
Internet of industrial smart things [6]	×	×	×	×	✓	×
Internet of healthcare [6]	✓	×	×	×	×	×
Mobile-as-a-service [38]	✓	×	×	✓	×	×
Mobile-as-manufacturing [38]	×	✓	×	×	✓	✓
Fine medicine [38]	✓	×	×	×	×	×
Intelligent disaster prediction [38]	✓	✓	×	×	×	✓
Extreme capacity backhaul [34]	✓	✓	✓	✓	✓	✓
Smart rail mobility [34]	×	×	×	×	×	✓
Connectivity in remote areas [17], [34]	✓	✓	✓	✓	✓	✓
Massive scale communications [39]	✓	✓	✓	✓	✓	✓

TABLE 2. 6G spectrum-level technology enablers along with their KPI impacts in system capacity, system latency, and system management: ✓and × symbolize yes and no, respectively, to a given KPI impact.

	KPI Impacts		
Proposed 6G Technologies	**System Capacity**	**System Latency**	**Spectrum Management**
An all-spectrum reconfigurable front-end for dynamic spectrum access [14]	✓	×	×
Superfast wireless broadband connectivity [40]	✓	✓	×
Above 6 GHz for 6G: from small cells to tiny cells [3]	✓	×	×
Transceivers with integrated frequency bands [3]	✓	×	×
Optical camera communication [21]	✓	✓	✓
Optical wireless communications [33]	✓	✓	✓
Multi-band ultrafast-speed transmission [41]	✓	✓	×
The hollistic management of communication, computation, caching, and control resources [42]	✓	✓	✓

[17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. As it is envisioned nowadays, 6G is driven by various anticipated applications such as multi-sensory XR applications [3], haptic communication [30], human bond communications [15], flying vehicles [21], and the Metaverse [31], [32], to mention a few (see Table 1). These applications will be realized using a combination of many promising 6G technology enablers. These enablers can be grouped as *6G spectrum-level enablers* (see Table 2), *6G infrastructure-level enablers* (see Table 3), and *6G algorithm/protocol-level enablers* (see Table 4) [33], [34]. Per these enablers, realizing 6G needs both an evolutionary and a revolutionary paradigm shift [3].

A revolutionary paradigm shift has to reckon with the following *fundamental 6G challenges* [43]:

1) *Guaranteeing an ultra-high data rate for most users.*
2) *Ensuring ultra-reliability and low latency for most users.*
3) *Managing ultra-heterogeneity.*
4) *Addressing ultra-high mobility.*

TABLE 3. 6G infrastructure-level technology enablers along with their KPI impacts in system capacity, system latency, and system management: ✓ and × symbolize yes and no, respectively, to a given KPI impact.

	KPI Impacts		
Proposed 6G Technologies	**System Capacity**	**System Latency**	**Spectrum Management**
Zero-touch network and service management [6]	×	✓	×
Three-dimensional (3D) coverage [42]	✓	✓	✓
Satellite-assisted IoT communications [44]	✓	×	✓
Supermassive multi-input multi-output (MIMO), large intelligent surfaces, and holographic beamforming [45]	✓ ×	✓ ×	× ×
Communication with reconfigurable intelligent surface (RIS) [3], [46], [47]	✓	✓	×
Airplane-aided integrated networking [48]	✓	×	✓
Tiny-cell communications and cell-free communications [49]	✓	×	×
Extremely large aperture arrays [50]	✓	×	×
Holographic massive MIMO [50]	✓	×	×
Six-dimensional positioning [50]	×	×	✓
Large-scale MIMO radar [50]	✓	×	×
Intelligent massive MIMO [50]	✓	✓	✓
Multi-purpose converged, full-spectral, and all-photonic radio access networks [38]	✓	✓	✓
Cell-free networks [35]	✓	×	×
Metamaterials-based antennas [35]	✓	✓	×
Fluid antenna [35]	✓	✓	×
Software-defined materials [35]	✓	✓	×
Programmable metasurfaces [35]	✓	✓	×
Wireless power transfer and energy harvesting [35]	×	×	✓
Hyperspectral space-terrestrial integration network [38]	✓	×	✓
Integrated access and backhaul networks [33], [34]	✓	×	✓
Antenna-on-glass [20]	✓	×	×
Internet of space things [14]	✓	×	✓
Super flexible integrated network [41]	✓	×	✓
Holographic radio and photodiode-coupled antenna arrays [38]	✓	×	×
Integrated terrestrial, airborne, and satellite networks [3]	✓	✓	✓
Energy transfer and harvesting [3]	×	×	✓

5) *Taming ultra-high complexity in 6G networks.*
6) *Being considerably energy efficient.*
7) *Enabling energy-efficient AI.*
8) *Incorporating various key performance indicators (KPIs) in the overall design.*
9) *Accommodating users' needs or perspectives.*
10) *Coping with the inevitable technological uncertainty associated with 6G technology enablers.*
11) *Ensuring security, privacy, and trust.*
12) *Attaining full intelligence and autonomy.*

Tackling the enumerated challenges amounts to surmounting a myriad of interdisciplinary, multidisciplinary, and transdisciplinary (IMT) challenges interwoven with many technological uncertainties and challenges [43], as elaborated in Section II-A. We now move on to our motivation.

B. MOTIVATION

To ameliorate the above-discussed fundamental challenges, one may undertake the design of 6G systems and networks while aiming at the simultaneous minimization of bandwidth consumption, power usage, and transmission delay, which can be achieved by minimizing the transmission of *semantically-redundant/irrelevant information* [43], [60], [61], [67], [68], [69]. This communication paradigm is widely recognized as *semantic communication* (SemCom) – envisioned first by Weaver around 1949 [70, Ch. 1] – and calls for an efficient transmission of *semantics* by a semantic transmitter followed by faithful recovery by a semantic receiver [61], [68], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80]. To this end, a DL-enabled SemCom has arisen as a promising 6G algorithm/protocol-level enabler via its inherent ability to minimize power usage, bandwidth consumption, and transmission delay while minimizing the transmission of semantically-irrelevant information [43], [60], [61], [67], [69].

A DL-enabled SemCom's considerable reduction in bandwidth consumption epitomizes a paradigm shift from "how to transmit" to "what to transmit" [81]. Regarding the latter, it is worth underscoring that conventional wireless communication systems have consistently approached the Shannon limit [59]. Therefore, one should originate a breakthrough that supports the insatiable desire for high data rates, the unparalleled proliferation of mobile devices, and the emergence of new and highly heterogeneous 6G applications and use cases [82]. In this vein, a DL-enabled SemCom is arising as a promising communication paradigm for the design, analysis, and optimization of 6G systems and 6G networks. As a *possibly revolutionary* communication paradigm, a DL-enabled SemCom can likely change the *status quo* – as traditional communication systems' designers have envisioned it [83] – that wireless connectivity is an opaque data pipe carrying messages whose context-dependent *meaning and effectiveness* have been

TABLE 4. 6G algorithm/protocol-level technology enablers along with their KPI impacts in system capacity, system latency, and system management: ✓ and × symbolize yes and no, respectively, to a given KPI impact.

Proposed 6G Technologies	KPI Impacts		
	System Capacity	System Latency	Spectrum Management
Pervasive AI [35]	✓	✓	✓
Edge AI [3], [5], [51]–[53]	✓	✓	✓
AI and photonics-based cognitive radio [38]	✓	✓	✓
Model-aided wireless AI [54], [55]	✓	✓	✓
Autonomous wireless systems with AI [56]	✓	✓	✓
AI/ML-driven air interface design and optimization [20]	✓	✓	✓
Networking with the sixth sense [20]	✓	✓	✓
Intelligent radio [57]	✓	✓	✓
AI-enabled closed-loop optimization [57]	✓	✓	✓
Intelligent wireless communication [57]	✓	✓	✓
Hardware-aware communication [57]	✓	✓	✓
Ultra-low-latency communication [40]	×	✓	✓
Data-oriented transmission [58]	✓	✓	✓
Semantic communication (SemCom) [42], [59]–[61]	✓	✓	✓
Tactile internet [21]	×	✓	✓
Multi-access edge computing [21]	✓	✓	✓
Intelligent internet of intelligent things [62]	✓	✓	✓
Network harmonization and interoperability [49]	×	×	✓
Intelligent proactive caching and mobile edge computing [49]	✓	✓	✓
Multi-objective optimization and routing optimization [49]	✓	✓	✓
Massive IoT and big data analytics [49]	✓	✓	✓
Configurable multi-antenna systems [49]	✓	×	×
Intelligent cognitive radio and self-sustaining wireless networks [49]	✓	×	×
Ambient backscatter communication [14]	✓	×	×
Self-driving networks [14]	×	✓	✓
Intelligent device-to-device communications [63]	✓	✓	✓
Device-centric wireless communications [64]	✓	✓	✓
Demand-driven opportunistic networking [64]	✓	✓	✓
Symbiotic radio [44]	✓	×	×
Super IoT [44]	✓	×	✓
The seamless integration of wireless information and energy transfer [19]	×	×	✓
Delta-orthogonal multiple access [65]	✓	×	×
Coded caching and rate splitting [34]	✓	✓	✓
Index modulation [66]	✓	×	×
Multi-mode multi-domain joint transmission [41]	✓	×	✓
Intelligent transmission [41]	✓	✓	✓
Orbital angular momentum (OAM) multiplexing [45]	✓	×	×
Blockchain-based spectrum sensing [45]	✓	×	×
Disaggregation and virtualization [39]	✓	×	✓
Ubiquitous sensing [40]	×	×	✓
Battery-free communication [42]	×	×	✓
Molecular communications and IoT [45]	✓	×	✓
Big data analytics for 6G [57]	✓	✓	✓
Compressed sensing and sparse coding [36]	×	✓	×
Edge caching and social networks [36]	✓	✓	✓
Quantum computing, communications, and networking [35], [45]	✓	✓	✓

ignored [83], [84]. In view of effectiveness, *goal-oriented SemCom*[1] focuses on employing semantic information – at a suitable time – efficiently for successful task execution, unlike SemCom which emphasizes semantic transmission for delivering the meaning behind the transmitted content [77].

SemCom deals with the transmission of complex data structures (e.g., patterns, features, and data lying on manifolds) or, in general, abstract concepts [85]. SemCom, where the effectiveness of semantic transmission is explicitly defined and focused on, can be qualified as a goal-oriented SemCom [85], [86]. Hence, SemCom is a broader concept than goal-oriented SemCom because the semantics of information are not necessarily linked to a system's overarching goal [85]. Per this view, goal-oriented SemCom is a subset of SemCom that takes a pragmatic approach to SemCom, where the receiver is interested in the semantics of the source's transmitted message and the message's effectiveness in accomplishing a certain goal(s) [85]. To this end, goal-oriented SemCom is aimed at extracting and transmitting only task-relevant information so that the transmitted signal

[1] Throughout this survey, task-oriented communication and goal-oriented communication are mainly presented under the heading goal-oriented SemCom. Nevertheless, the authors of [77] underline that goal communication is much broader than SemCom, which they classify – according to Weaver's vision – as *semantic-level SemCom* and *effectiveness-level SemCom*.

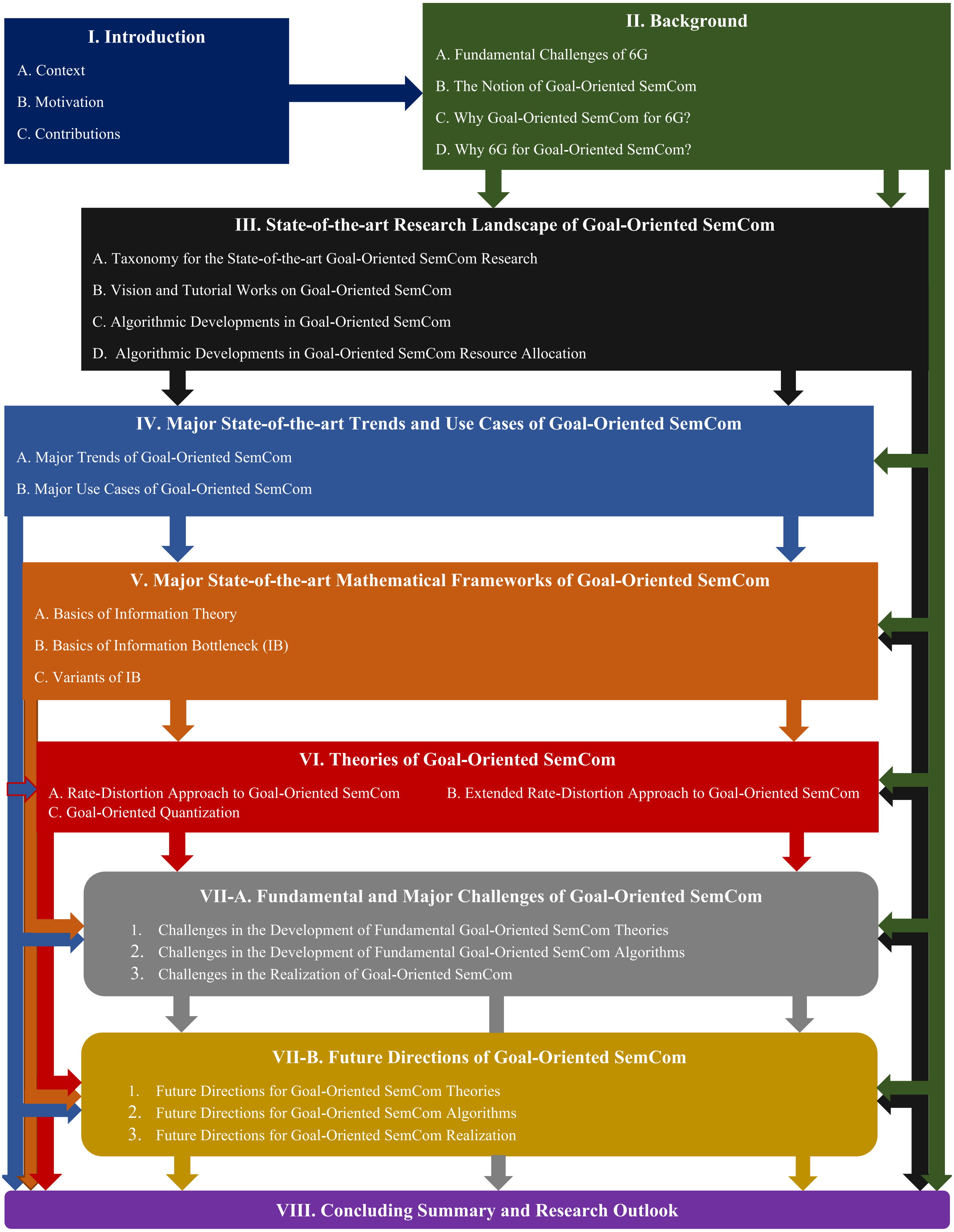

I. Introduction
A. Context
B. Motivation
C. Contributions
II. Background
A. Fundamental Challenges of 6G
B. The Notion of Goal-Oriented SemCom
C. Why Goal-Oriented SemCom for 6G?
D. Why 6G for Goal-Oriented SemCom?
III. State-of-the-art Research Landscape of Goal-Oriented SemCom
A. Taxonomy for the State-of-the-art Goal-Oriented SemCom Research
B. Vision and Tutorial Works on Goal-Oriented SemCom
C. Algorithmic Developments in Goal-Oriented SemCom
D. Algorithmic Developments in Goal-Oriented SemCom Resource Allocation
IV. Major State-of-the-art Trends and Use Cases of Goal-Oriented SemCom
A. Major Trends of Goal-Oriented SemCom
B. Major Use Cases of Goal-Oriented SemCom
V. Major State-of-the-art Mathematical Frameworks of Goal-Oriented SemCom
A. Basics of Information Theory
B. Basics of Information Bottleneck (IB)
C. Variants of IB
VI. Theories of Goal-Oriented SemCom
A. Rate-Distortion Approach to Goal-Oriented SemCom
B. Extended Rate-Distortion Approach to Goal-Oriented SemCom
C. Goal-Oriented Quantization
VII-A. Fundamental and Major Challenges of Goal-Oriented SemCom
1. Challenges in the Development of Fundamental Goal-Oriented SemCom Theories
2. Challenges in the Development of Fundamental Goal-Oriented SemCom Algorithms
3. Challenges in the Realization of Goal-Oriented SemCom
VII-B. Future Directions of Goal-Oriented SemCom
1. Future Directions for Goal-Oriented SemCom Theories
2. Future Directions for Goal-Oriented SemCom Algorithms
3. Future Directions for Goal-Oriented SemCom Realization
VIII. Concluding Summary and Research Outlook

FIGURE 1. Concept map and structure of this survey paper.

is substantially compressed, communication efficiency is improved, and low end-to-end latency is achieved [87].

Since a (DL-enabled) SemCom is a revolutionary communication paradigm for 6G, goal-oriented SemCom embodies a promising paradigm shift for the design, analysis, and optimization of 6G systems and 6G networks. On the other hand, 6G is vital for the realization of major goal-oriented SemCom use cases such as autonomous transportation, consumer robotics, environmental monitoring, telehealth, smart factories, and networked control systems (NCSs). These paradigms of *6G for goal-oriented SemCom* and *goal-oriented SemCom for 6G* demand a tighter integration of 6G and goal-oriented SemCom. In facilitating this purpose, this survey leads to the contributions listed in Section I-C.

C. CONTRIBUTIONS

This survey paper provides a *comprehensive treatment* of goal-oriented SemCom, except its performance assessment metrics detailed by our survey in [43]. Particularly, the key contributions of this survey article are itemized below.

1) It first discusses the fundamental challenges of 6G.
2) It then provides a taxonomy for the state-of-the-art goal-oriented SemCom research.
3) It then details the state-of-the-art research landscape of goal-oriented SemCom.
4) It then presents the major state-of-the-art trends and use cases of goal-oriented SemCom.
5) It then outlines state-of-the-art mathematical frameworks of goal-oriented SemCom.
6) It then discusses state-of-the-art theories of goal-oriented SemCom.
7) It then exposes the fundamental and major challenges of goal-oriented SemCom theories, algorithms, and realization.
8) Finally, it offers promising future directions for goal-oriented SemCom theories, algorithms, and realization.

To put these contributions in perspective, Table 5 compares and contrasts them with the scope of state-of-the-art survey/tutorial papers. Meanwhile, the concept map and structure of this survey are schematized in Fig. 1, which informs this paper's organization. Section II presents

TABLE 5. Scope of this survey paper with respect to (w.r.t.) the scope of prior survey/tutorial papers on goal-oriented SemCom. "Ref." abbreviates Reference.

Ref.	Key focus of the survey/tutorial in the reference	How our survey paper differs from the survey/tutorial in the reference
[60]	It outlines the 6G visions and enablers for SemCom; provides detailed overview of SemCom developments; reviews semantic- and communication-related challenges and techniques for 6G SemCom; and outlines future directions for SemCom research. Although this is a SemCom-oriented paper, it also provides a few highlights on goal-oriented SemCom techniques, trends, use cases, and theories.	It presents the fundamental challenges of 6G and the state-of-the-art mathematical frameworks of goal-oriented SemCom; details the state-of-the-art goal-oriented SemCom techniques, major trends, major use cases, and theories; presents the fundamental and major challenges of goal-oriented SemCom; and offers future directions for goal-oriented SemCom theories, algorithms, and realization.
[75]	It provides an overview on SemCom by briefly discussing SemCom with theory, frameworks, and DL-enabled system design. It also discusses some performance metrics for SemCom, and offers open questions for SemCom research. Even though this is a SemCom-oriented tutorial, it highlights a few goal-oriented SemCom techniques and frameworks.	It presents the fundamental challenges of 6G, the state-of-the-art goal-oriented SemCom trends and use cases, goal-oriented SemCom theories, and the fundamental and major challenges of goal-oriented SemCom; offers future directions for goal-oriented SemCom theories, algorithms, and realization; and details the state-of-the-art goal-oriented SemCom techniques and frameworks.
[85]	This mini-survey presents how problems of IoT networks such as data compression, data clustering, data estimation, and ML are related to goal-oriented communication; highlights a few challenges of goal-oriented communication	It presents the fundamental challenges of 6G; details the state-of-the-art goal-oriented SemCom techniques, trends, use cases, mathematical frameworks, and theories; exposes the fundamental and major challenges of goal-oriented SemCom; offers future directions for goal-oriented SemCom theories, algorithms, and realization.
[77]	It presents the evolution of AI-empowered SemCom, semantic information theory, semantic metrics, techniques of semantic level- and effectiveness level-SemCom, and challenges and open issues in semantics-empowered communication.	It presents the fundamental challenges of 6G; details the state-of-the-art goal-oriented SemCom trends, use cases, and theories; provides detailed information on the state-of-the-art goal-oriented SemCom techniques, frameworks, and challenges (both fundamental and major); offers future directions for goal-oriented SemCom theories, algorithms, and realization.
[78]	It presents an overview of the information theoretical foundations of semantic- and task-oriented communications; details practical data-driven approaches to SemCom of various information sources including image, text and video; reviews DL-aided approaches to communicating practical information sources over constrained communication channels; offers introduction to tools and advancements in semantic and task-oriented communications.	It presents the fundamental challenges of 6G; details the state-of-the-art techniques, trends, use cases, and theories of goal-oriented SemCom; exposes the major and fundamental challenges of goal-oriented SemCom; offers future research directions for goal-oriented SemCom theories, algorithms, and realization.
[88]	Focusing on cyber-physical systems (CPS), this survey presents literature review on the theoretical perspectives of information/communication theory, control theory, and computer science; devise a generic framework for designing a task-oriented system and adapt it for several 6G use cases, which they have also presented a literature review for; discuss the challenges and open problems of its proposed task-oriented communications design.	It presents the fundamental challenges of 6G; details the state-of-the-art goal-oriented SemCom techniques, trends, and theories; exposes the fundamental and major challenges of goal-oriented SemCom; offers future directions for goal-oriented SemCom theories, algorithms, and realization.

TABLE 6. List of symbols.

Notation	Meaning
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of non-negative real numbers
$\mathbb{R}^n$	The set of n-dimensional vectors of real numbers
$\mathbb{R}^n_+$	The set of n-dimensional vectors of non-negative real numbers
$\mathbb{R}^{m\times n}$	The set of $m \times n$ matrices of real numbers
$:=$	Equality by definition
$\sim$	Distributed as
$\\|\cdot\\|$	The Euclidean norm
$(\cdot)^T$	Transpose
$\mathcal{O}(\cdot)$	The Landau notation
$[n]$	$\{1, 2, \ldots, n\}, n \in \mathbb{N}$
$\mathbb{N}_{>k}$	$\{k, k+1, k+2, \ldots\}, k \in \mathbb{N}$
$\min(\cdot)$ $(\min\{\cdot\})$	Minimum
$\max(\cdot)$ $(\max\{\cdot\})$	Maximum
$\mathbb{E}\{\cdot\}$	Expectation
$\mathbb{E}_{p(x)}\{\cdot\}$	Expectation w.r.t. a distribution function $p(x)$
$\mathbb{P}(\cdot)$	Probability

background information. Section III details the state-of-the-art research landscape of goal-oriented SemCom. Section IV presents the major state-of-the-art trends and use cases of goal-oriented SemCom. Section V outlines state-of-the-art mathematical frameworks of goal-oriented SemCom. Section VI discusses state-of-the-art theories of goal-oriented SemCom. Section VII exposes and offers the challenges and future research directions, respectively, of goal-oriented SemCom. Section VIII provides a concluding summary and research outlook.

Notation: Scalars, vectors, and matrices are represented by italic lowercase letters, bold lowercase letters, and bold uppercase letters, respectively. Random variables (RVs) are denoted by italic uppercase letters. Sets, quantizers, and quantization regions are designated by calligraphic letters. The symbols used in this paper are defined in Table 6.

II. BACKGROUND

To offer a top-down background information for this survey article, this section details the fundamental challenges of 6G, the notion of goal-oriented SemCom, why goal-oriented SemCom is for 6G, and why 6G is for goal-oriented SemCom. We start with the fundamental challenges of 6G.

A. FUNDAMENTAL CHALLENGES OF 6G

The 12 fundamental challenges of 6G are henceforth elaborated, beginning with guaranteeing an ultra-high data rate for most users.

1) CHALLENGE I: GUARANTEEING AN ULTRA-HIGH DATA RATE FOR MOST USERS

To guarantee an ultra-high data rate for most users who may wish to use 6G applications such as multi-sensory XR applications and holography, 6G needs to provide ultra-high data rates of up to 1 Tb/s. To do so successfully, the design of 6G networks must exploit Terahertz (THz) and visible light frequency bands. However, realizing both THz communication and visible light communication (VLC) in 6G faces numerous major challenges. To begin with sub-THz or THz communication, the following are its major challenges: radio frequency (RF) architecture and signal processing optimization [34]; RF impairments and phase noise [34]; increasing the communication distance [34]; realizing band-limited converters [34]; THz channel modeling for spatial consistency [27]; reducing computational complexity in massive antenna spatial multiplexing and beam codebooks [27]; deafness problem and line-of-sight blockage [14]; the propagation delay across the antenna array is comparable to the symbol duration [20]; and robust THz channel modeling that accurately captures propagation effects (reflection, scattering, diffraction), concerning different materials, as well as atmospheric attenuation and molecular absorption (at various transmission windows) [14]. To enable the materialization of VLC, major challenges are in order: field-of-view alignment and shadowing [89]; VLC channel modeling [34]; receiver design and energy efficiency [89]; light emitting diode (LED) to Internet connectivity [89]; inter-cell interference, uplink and RF augmentation [89]; LEDs' limited modulation bandwidth and slow modulation response [28]; non-linearity compensation for orthogonal frequency division multiplexing (OFDM)-based VLC [28]; and interference mitigation for VLC MIMO [28].

2) CHALLENGE II: ENSURING ULTRA-RELIABILITY AND LOW LATENCY FOR MOST USERS

A number of envisioned 6G applications, such as industrial automation and holography, require an ultra-low latency in the order of a fraction of a millisecond [3]. To properly support such 6G applications, 6G networks must enable various 6G services with *very high reliability and low latency anytime and anywhere*. However, the following fundamental problems arise in the pursuit of URLLC [90]: 3D performance analysis of rate-reliability-latency (R^{2}L) targets, characterization of R^{2}L targets, quantification of achievable R^{2}L targets, and fundamental R^{2}L limits [3].

3) CHALLENGE III: MANAGING ULTRA-HETEROGENEITY

Regarding the management of ultra-heterogeneity in light of highly heterogeneous 6G driving applications (see Table 1), the considerably different requirements of the 6G driving applications must be satisfied by intelligent, integrated, and high-dimensional 6G networks for the applications to work seamlessly. Such ultra-heterogeneity management may hinge on the successful integration of communication, control, computing, localization, and sensing (3CLS). Nonetheless, realizing 3CLS faces some open problems, such as the design of 3CLS metrics [3], joint 3CLS optimization [3], application-specific design of an intelligent holistic orchestration platform to coordinate all network resources [21], and artificial intelligence (AI)-enabled 3CLS [3].

4) CHALLENGE IV: ADDRESSING ULTRA-HIGH MOBILITY

For addressing ultra-high mobility, 6G is expected to offer consistent service experiences to many users in ultra-high-speed trains, flying taxis, self-driving cars, and supersonic aeroplanes. To overcome the considerable Doppler spread caused by the ultra-high mobility of these users, Doppler spread-resistant and robust transceivers will be needed. Nonetheless, designing such transceivers for millimeter wave (mmWave) and THz wireless communication systems is a fundamental challenge.

5) CHALLENGE V: TAMING ULTRA-HIGH COMPLEXITY IN 6G NETWORKS

Apart from addressing ultra-mobility, taming ultra-high complexity in 6G networks is also a fundamental challenge. For seamless operation and connectivity, 6G networks must incorporate integrated multi-tier networks such as *space-air-ground integrated networks* (SAGINs) [94], [95]. SAGIN entails numerous substantial challenges, such as optimal routing, scheduling, and resource allocation of multi-tier networks [94], [95]. Aside from SAGIN, 6G networks can also comprise floating, flying, and submerging base stations, thereby forming a number of 3D multi-tier networks. Nonetheless, taming ultra-high complexity in 3D multi-tier networks is quite challenging due to the following open problems: 3D performance metrics, 3D propagation modeling, 3D mobility management, and 3D network optimization [3].

6) CHALLENGES VI AND VII: BEING CONSIDERABLY ENERGY EFFICIENT AND ENABLING ENERGY-EFFICIENT AI

Aside from the ultra-complexity, ultra-mobility, and ultra-heterogeneity management in 6G networks, achieving energy efficiency is also a fundamental challenge. To this end, being considerably energy efficient across the 6G networks and enabling energy-efficient AI are fundamental challenges. Regarding the former, 6G networks are expected to serve 10^7 devices/km^2 [45]. This entails a fundamental energy efficiency challenge – not only for device design, but also for network design, provisioning, and optimization – that the overall network energy consumption and hardware energy consumption must be minimal. Proceeding to the fundamental challenge of enabling an energy-efficient AI, the amount of power used up by AI for training and frequent fine-tuning is ginormous. Accordingly, should 6G networks be intelligent and energy efficient, an energy-efficient AI is needed. To enable such AI, *neuromorphic computing* [96], [97], [98], [99] is promising. Nevertheless, numerous lines of multidisciplinary research are needed in applications, algorithms, software, devices, and materials [100].

7) CHALLENGE VIII: INCORPORATING VARIOUS KPIS IN THE OVERALL DESIGN

Apart from energy efficiency and management of ultra-complexity, ultra-mobility, and ultra-heterogeneity, 6G must be fundamentally designed w.r.t. various KPIs. This is because there are numerous heterogeneous 6G driving applications (see Table 1) that will impose stringent requirements and, consequently, different KPIs on the 6G network.

8) CHALLENGE IX: ACCOMMODATING USERS' NEEDS OR PERSPECTIVES

Besides designing w.r.t. various KPIs, the fundamental design of 6G and 6G networks must also accommodate the users' perspectives or needs. As advocated by the authors of [101], the users' needs should be rigorously addressed in the design, analysis, optimization, and standardization of 6G communication systems. To this end, high-fidelity voice and normal voice with ultra-low power consumption; long range, low bit rate, ultra-low power, and longer latency; normal range, low bit rate, and medium delay are crucial 6G service categories [101].

9) CHALLENGE X: COPING WITH THE INEVITABLE TECHNOLOGICAL UNCERTAINTY ASSOCIATED WITH 6G TECHNOLOGY ENABLERS

In connection with the design of 6G and 6G networks, an important fundamental problem is coping with the inevitable technological uncertainty associated with 6G technology enablers [102]. This is required for the fact that there is a significantly high uncertainty inextricably linked to key technology enablers that can have a high impact on 6G [102]. Some examples of such enablers are neural interfaces, complementary metal oxide semiconductors (CMOS) in THz, all-day wearable displays, lossless materials at THz, and user apps running on the edge [102].

10) CHALLENGE XI: ENSURING SECURITY, PRIVACY, AND TRUST

Apart from 6G design, management, and optimization, a pivotal and overarching fundamental challenge is ensuring security, privacy, and trust. To guarantee privacy and security to every 6G user, trust should be embedded into 6G networks [92], [102]. Meanwhile, the IMT issues of 6G security, trust, and privacy boil down to six interwoven challenges [92]:

- *End-to-end embedded trust in 6G*
- *Novel threats at the 6G scale*
- *Machine learning (ML) as a tool and implicit risk for 6G*
- *Post-quantum cryptography and security architecture for 6G*
- *Physical layer (PHY) security*
- *Exploitation of privacy as a resource in 6G*

In light of the itemized challenges, the 6G security, privacy, and trust are confronted with an astronomical number of challenges and open problems, as listed in Table 7).

11) CHALLENGE XII: ATTAINING FULL INTELLIGENCE AND AUTONOMY

Attaining full intelligence and autonomy is of paramount importance since the 6G network management, 6G design,

TABLE 7. Challenges and open problems for the 6G security, privacy, and trust.

Themes of 6G security, privacy, and trust	Challenges and open problems
6G security and privacy	★ Deploying 6G security and privacy measures in 6G architecture, which is not yet defined [6]. ★ Security on-demand for 6G applications [6]. ★ Context-aware security protection [91]. ★ IMT techniques that guarantee the confidentiality, integrity, authenticity, and availability of information in 6G radio access network (6GRAN) and 6G core network. ★ Future fate of SIM cards [92]. ★ Use of asymmetric cryptography [92]. ★ Physical layer security (PHYSec) in a challenged 6G environment (large network scalability, several heterogeneous devices, and many forms of malicious attacks) [92]. ★ A stand-alone 6G PHYSec versus 6G PHYSec operating with upper layers [92]. ★ Novel differential privacy approaches [92].
6G trust	★ Robust 6G trust networking [92]. ★ Robust 6G embedded trust [92]. ★ 6G trust modeling, policies, and mechanisms [92].
AI/ML-enabled 6G security and privacy	★ What if AI tools are used for attacking purpose? [6] ★ Ensuring security in AI algorithms [6]. ★ Designing and deploying privacy-enabled AI model [6]. ★ Novel AI/ML-based proactive security [93]. ★ Defending against ML orchestrated attacks [92]. ★ Deep learning (DL)-based packet/byte-level 6G security [92].

6G network optimization, 6G network orchestration, 6G resource allocation, and 6G security, trust, and privacy hinge on full intelligence and autonomy, should it be achieved. Propelled by the remarkable advancements of the various DL techniques [103], [104], [105], [106], [107] that can effectively learn a stationary distribution from a dataset, there has been a persistent push by the multidisciplinary AI/ML [108], [109], [110], [111], [112] research communities to learn *non-stationary distributions* on a continuous basis. Learning and reasoning from non-stationary distributions on a continuous basis can pave the way to full intelligence. Full intelligence is often referred to under the name *artificial general intelligence* (AGI) [113], [114], [115], [116], [117], and considerable research endeavors have continued to be made toward the realization of AGI [118], [119], [120], [121], [122] (despite some renewed skepticism [123]). Meanwhile, some contemporary researchers believe that a combination of lifelong ML/DL [124], [125], [126], [127], [128], [129], [130], [131][2]; meta-learning [132], [133], [134]; probabilistic ML [135]; causal inference [136], [137], [138]; causal representation learning [139], [140]; causal ML [141]; commonsense reasoning [142], [143]; neural-symbolic learning and reasoning (neurosymbolic AI) [144], [145], [146], [147]; and knowledge representation and reasoning [148], [149], [150], [151], [152] can pave the way to AGI. On the other hand, the concept of full autonomy is much more complicated than AGI alone, as autonomous systems have to deal with many *unknown unknowns* [108], [153]. Thus, autonomous systems require, among other things, *sensing*, *perception*, *knowledge repository*, *self-adaptation*, *reflection*, *goal management*, and *planning* in a never-ending manner [154], [155].

All the aforementioned fundamental challenges of 6G are summarized in Table 8. We now move on to our discussion on the essence of goal-oriented SemCom.

B. THE NOTION OF GOAL-ORIENTED SEMCOM

Even though Shannon purposely ignored[3] the semantic aspects of communication in his classic work documented by [156], the notion of SemCom was first revealed by Weaver [70, Ch. 1]. Specifically, Weaver had envisaged communication using semantics and outlined three levels of communication – schematized in Fig. 2 – that can fundamentally differentiate the broad subject of communication [70, p. 4]:

- "**Level A**. How accurately can the symbols of communication be transmitted? (The technical problem).
- '**Level B**. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem).
- '**Level C**. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem)."

As depicted in Fig. 2, Weaver's three levels of communication are *technical level* (Level A), *semantic level* (Level B), and *effectiveness level* (Level C). These levels of communication correspond to the *technical problem*, the *semantic problem*, and the *effectiveness problem*.

Driven by Shannon's information theory, which has steered the design of generations of communication systems, the technical problem revolves around the accurate transmission of a transmitted message's symbols/bits. This old paradigm renders wireless connectivity to be an opaque data pipe carrying messages whose context-dependent meaning and effectiveness have been neglected [83] while the sender and

[2] Lifelong learning is also known as continual learning, class incremental learning, and never-ending learning.

[3] "Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem" [156, p. 1].

TABLE 8. Summary on the 12 fundamental challenges of 6G.

Fundamental 6G challenges	Reasons and remarks
I – Guaranteeing an ultra-high data rate for most users	(sub-)THz communication and VLC have the potential to guarantee an ultra-high data rate for most users. Nonetheless, (sub-)THz communication faces the following challenges: ★ RF architecture and signal processing optimization [34]; RF impairments and phase noise [34]; increasing the communication distance [34]; realizing band-limited converters [34]; THz channel modeling for spatial consistency [27]; reducing computational complexity in massive antenna spatial multiplexing and beam codebooks [27]; deafness problem and line-of-sight blockage [14]; the propagation delay across the antenna array is comparable to the symbol duration [20]; and robust THz channel modeling that accurately captures propagation effects, atmospheric attenuation, and molecular absorption [14]. ★ VLC also faces the following challenges: Field-of-view alignment and shadowing [89]; VLC channel modeling [34]; receiver design and energy efficiency [89]; LED to Internet connectivity [89]; inter-cell interference, uplink and RF augmentation [89]; LEDs' limited modulation bandwidth and slow modulation response [28]; non-linearity compensation for OFDM-based VLC [28]; and interference mitigation for VLC MIMO [28].
II – Ensuring ultra-reliability and low latency for most users	In the pursuit of URLLC, the following fundamental problems arise [90]: ★ 3D performance analysis and characterization of R^2L targets, quantification of achievable R^2L targets, and fundamental R^2L limits [3].
III – Managing ultra-heterogeneity	★ Although ultra-heterogeneity management may hinge on the successful integration of 3CLS, realizing 3CLS faces the following open problems: The design of 3CLS metrics [3]; joint 3CLS optimization [3]; application-specific design of an intelligent holistic orchestration platform to coordinate all network resources [21]; and AI-enabled 3CLS [3].
IV – Addressing ultra-high mobility	★ Even though Doppler spread-resistant and robust transceivers will be needed to overcome the considerable Doppler spread caused by the ultra-high mobility of 6G users, designing such transceivers for mmWave and THz wireless communication systems is a fundamental challenge.
V – Taming ultra-high complexity in 6G networks	★ 6G networks can comprise several 3D multi-tier networks, whose ultra-high complexity will be quite challenging to tame due to the following open problems: 3D performance metrics, 3D propagation modeling, 3D mobility management, and 3D network optimization [3].
VI and VII – Being considerably energy efficient and enabling energy-efficient AI	★ Should 6G networks be intelligent and energy efficient, one must devise an energy-efficient AI, which can potentially be realized using neuromorphic computing [96]–[99]. Nevertheless, numerous lines of multidisciplinary research are needed in applications, algorithms, software, devices, and materials [100].
VIII – Incorporating various KPIs in the overall design	★ 6G must be fundamentally designed w.r.t. various KPIs for there are numerous heterogeneous 6G driving applications that will impose stringent requirements and, thus, different KPIs on the 6G network.
IX – Accommodating users' needs or perspectives	★ The fundamental design of 6G and 6G networks must also accommodate the users' perspectives or needs: for instance, high-fidelity voice and normal voice with ultra-low power consumption; long range, low bit rate, ultra-low power, and longer latency; normal range, low bit rate, and medium delay [101].
X – Coping with the inevitable technological uncertainty associated with 6G technology enablers	★ There is a considerably high uncertainty inextricably linked to key technology enablers, that can have high impact on 6G, such as neural interfaces, CMOS in THz, all-day wearable displays, lossless materials at THz, and user apps running on the edge [102].
XI – Ensuring security, privacy, and trust	The IMT issues of 6G security, trust, and privacy boil down to six interwoven challenges [92]: ★ End-to-end embedded trust in 6G ★ Novel threats at the 6G scale. ★ ML as a tool and implicit risk for 6G. ★ Post-quantum cryptography and security architecture for 6G. ★ PHY security. ★ Exploitation of privacy as a resource in 6G.
XII – Attaining full intelligence and autonomy	★ According to some contemporary researchers, full intelligence or AGI is a fundamental challenge that can be realized via a combination of lifelong ML/DL [124]–[131]; meta-learning [132]–[134]; probabilistic ML [135]; causal inference [136]–[138]; causal representation learning [139], [140]; causal ML [141]; commonsense reasoning [143], [143]; neurosymbolic AI [144]–[147]; and knowledge representation and reasoning [148]–[152]. ★ Full autonomy is much more complicated than AGI alone, as autonomous systems have to deal with many unknown unknowns [108], [153], and hence require sensing, perception, knowledge repository, self-adaptation, reflection, goal management, and planning – on a never-ending manner. [154], [155].

receiver are presumed to be agents without intelligence [157]. As a result, large semantically-irrelevant/redundant data are transmitted whilst wasting scarce communication resources such as transmission power and bandwidth [157]. Apart from wasting bandwidth and power, acquiring, processing, and sending unnecessarily huge distributed real-time data – that is likely to be useless to the end users or outdated when they reach them – will result in increased latency, communication bottlenecks, and safety issues in cyber-physical systems and autonomous NCSs [83]. Thus, communication system designers must look beyond the technical problem for a paradigm where communication in itself is a means to attaining specific goals rather than an end goal [83], [158]. W.r.t. the attainment of specific goals, the semantic problem and effectiveness problem can come into play.

As seen in Fig. 2, the semantic problem is grounded on how the desired meaning is being conveyed precisely by the transmitted symbols. Such a semantic-centric viewpoint is the heart of SemCom, which stresses semantic transmission for data reduction and the delivery of the meaning behind the transmitted content [77]. As such, SemCom relies on a *knowledge base* (KB) that is shared between the source and

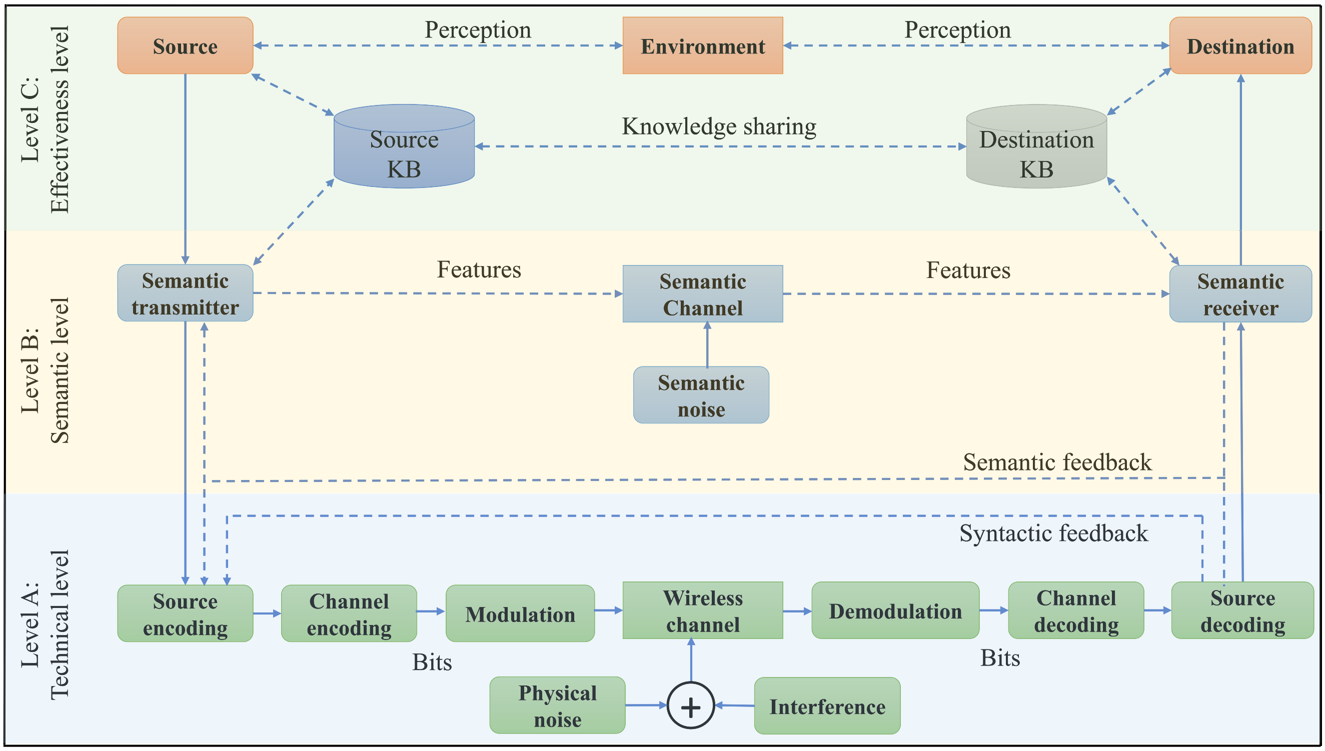


FIGURE 2. Weaver's three levels of communication – redrawn from [73, Fig. 1]. KB: knowledge base.

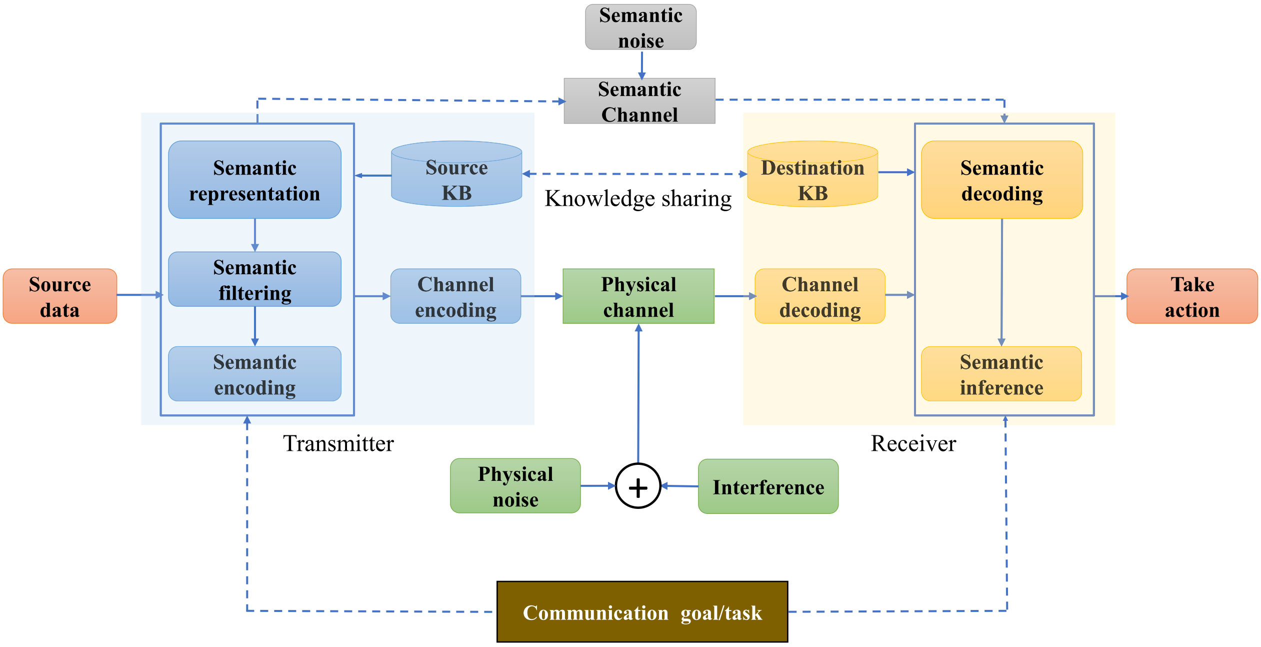


FIGURE 3. System model for goal-oriented SemCom – redrawn from [60, Fig. 6(c)].

the destination, secure *human-to-human* (H2H), *human-to-machine* (H2M), and *machine-to-machine* (M2M) systems can be designed using SemCom [76]. SemCom is also promising as a key enabler of *6G edge intelligence* with efficient computation and communication overheads while overcoming the challenges faced by 6G communication networks [60], [160]. Therefore, SemCom endows 6G with the possibility of designing 6G systems that can benefit significantly from a system design that incorporates not only a semantic component but also an effectiveness component.

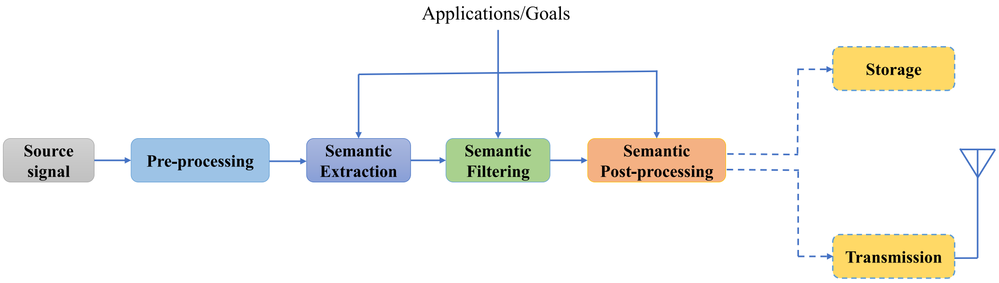


FIGURE 4. Goal-oriented semantic signal processing framework – redrawn from [159, Fig. 12], [75, Fig. 6].

As to the effectiveness component, the effectiveness problem (per Fig. 2) centers on how conduct in the desired way is effectively affected by the received meaning [70, p. 4]. This design paradigm is at the heart of goal-oriented SemCom, which concentrates on deploying semantic information effectively – at an appropriate time – for successful task execution [77]. In this vein, goal-oriented SemCom aims to enable interested communicating parties to achieve a joint communication goal/task [60], [158]. This brings us to our discussion concerning a goal-oriented SemCom system model that is depicted in Fig. 3.

To complete a joint communication goal/task, Fig. 3 illustrates a system model for goal-oriented SemCom. The effectiveness-level SemCom's transmitter transforms the source data into semantically encoded information via semantic representation, semantic filtering, and semantic encoding in a sequential process. This process is carried out using the source KB w.r.t. a given communication goal/task. W.r.t. a communication goal/task and a destination KB that largely share common knowledge with a source KB, the receiver aims to take a desired action by acting on the output of the channel decoder via semantic decoding followed by semantic inference. The inference module's output – for instance, in self-driving autonomous cars – includes action execution instructions for accelerating and braking; changing the angle for the steering wheel and flashing the headlights; and responding to pedestrians, roadblocks, and traffic signal changes, among other actions [60]. At the receiver, each of these goals requires (possibly application/goal-tailored) semantic extraction (SE) followed by semantic filtering and semantic post-processing prior to source signal transmission [159], as depicted in Fig. 4.

Fig. 4 shows the task/goal-oriented semantic signal processing framework put forward by the authors of [159]. These authors propose a framework that comprises pre-processing, SE,[4] semantic filtering, semantic post-processing, and storage/transmission in a sequence. When it comes to storage/transmission scheduling, the pre-processing block first transforms the input signal – following a likely pre-filtering to reduce noise and/or interference – into an appropriate domain for efficient component detection/classification [159]. The SE block employs this transformed input under a time-varying application/goal and generates the corresponding multi-graph description and attribute sets [159]. Thereafter, the semantic filtering block carries out semantic filtering per the local and time-varying goals to produce semantic data [159]. The goal-filtered semantic data are then fed to the semantic post-processing block (see Fig. 4). The semantic post-processing block finally schedules – while incorporating the (time-varying) local goals – either transmission or storage per the receiver's communication goals/tasks [159], [161]. In the context of this goal-oriented semantic signal processing framework, the chief signal processing problems encountered in the internet of things (IoT) networks, such as data compression, data clustering, data estimation, and ML, are related to the paradigm of goal-oriented SemCom [85].

Per goal-oriented SemCom's articulated notion, goal-oriented SemCom is useful for 6G, since communication is not an end but a means to achieve specific goals [83], leading to our discussion on why goal-oriented SemCom is for 6G.

C. WHY GOAL-ORIENTED SEMCOM FOR 6G?

The 6G spectrum-level enablers (Table 2), 6G infrastructure-level enablers (see Table 3), and 6G algorithm/protocol-level enablers (see Table 4) have inspired many *6G trends and use cases* [68]: Convergence of communications, computing, control, localization, and sensing [3]; intelligent distributed computing and data analytics, network densification, and use of mobile edge, cloud, and fog computing [40]; AI-enabled autonomous wireless networks, and convergence of intelligent sensing, communication, computing, caching, and control [45]; smart cars and smart manufacturing [38]; multi-sensory holographic teleportation, real-time remote healthcare, autonomous cyber-physical systems, intelligent industrial automation, high-performance precision agriculture, space connectivity, and smart infrastructure and environments [14]; digital twin worlds [20], [162], [163], [164], [165]; knowledge systems, ubiquitous universal computing, and man-machine interfaces [20]; *maglev transportation*, traveling by air and sea, intelligent driving [36], internet of vehicles, and space exploration [36]; intelligent vehicle-to-everything [93]; symbiotic autonomous systems [166]; collaborative robots [62]; ubiquitous super 3D connectivity [19]; application-aware data burst forwarding, emergency and disaster rescue, and socialized IoT [167]; time sensitive and time engineered applications [22]; massive twinning [23]; gadget-

[4] In goal-oriented SemCom, SE must necessarily capture pragmatic information [60], whereas in SemCom, it revolves around semantic information.

free communication [6]; intelligent internet of medical things [168]; globalized ubiquitous connectivity, enhanced on-board connectivity, and pervasive intelligence [10].

Regarding one or more of the aforementioned 6G trends and use cases, supporting the simultaneous requirements of the eMBB, URLLC, and mMTC services can be a formidable challenge because of the deluge of data that should be transmitted by the respective applications. Specifically, 6G use cases, such as autonomous transportation, consumer robotics, environmental monitoring, telehealth, smart factories, and NCSs,[5] demand very high reliability, ultra-low latency, and ultra-large transmission bandwidth. These stringent requirements can be met by transmitting only the information that is semantically relevant for the effective performance of the desired action. Consequently, goal-oriented SemCom is crucial for designing and realizing 6G w.r.t. minimum power usage, bandwidth consumption, and transmission delay, while aiming to effectively achieve one or more goals by underpinning the scalability of future networked intelligent systems [83]. Supporting the scalability of future massive networked intelligent systems, semantic-empowered communication will enhance network resource usage, energy consumption, and computational efficiency significantly, paving the way for the design of next-generation real-time data networking [83]. This type of semantic networking will make it possible to transmit only informative data samples and convey only the information that is relevant, useful, and valuable for achieving its defined goals [83]. Accordingly, goal-oriented SemCom is a prime enabler for emerging 6G use cases, such as massive IoT, massive *industrial IoT* (IIoT), and Metaverse, as highlighted below.

IoT systems enable a seamless integration of the interconnected physical world and its programmable digital representation, forming a *cyber-physical continuum* with advanced intelligence and ubiquitous connectivity [169]. Per this vision, it is indispensable to provide *distributed AI services* – with high reliability and ultra-low latency – for time-critical IoT applications, such as autonomous driving [169]. Influenced by emerging 6G use cases such as autonomous driving and XR, the current trend in IoT is marked by a growing demand for wider bandwidth and increased energy consumption [169]. Because of the limited spectrum, computing power, and energy, chasing these requirements will likely lead to a *performance bottleneck* [169]. Such a bottleneck can be surmounted by transmitting only semantically-relevant information that would inform the effective execution of a given task(s). Therefore, (goal-oriented) SemCom is essential for realizing the promises of massive 6G IoT systems.

Leveraging the availability of huge data produced by sensors and various devices, IIoT enhance the accuracy, efficiency, and reliability of an industrial manufacturing process, enabling access to a much more complicated view of the current state of the system [88]. Nevertheless, extracting useful information can be challenging, given the limitations in the communications rate in rural areas and/or the processing power of actuators and controllers [88]. What can also be a challenge for the manufacturing value chain is the numerous tiny elements producing several megabytes of data per second, making processing power and communication the bottlenecks of the IIoT system [88]. In overcoming such bottlenecks of the IIoT system, goal-oriented SemCom is crucial, since it extracts only meaningful data generated by any controller/actuator of the system for the effective execution of a task by the system.

As a network of interconnected virtual worlds that are generated by computers, the Metaverse needs an infrastructure for sensing, actuation, communication, networking, computation, and storage. The Metaverse infrastructure feeds big data to *the Metaverse engine* that is enabled by core 6G technologies for interactivity (e.g., XR and BCI), tactile internet, digital twin, AI, and the Metaverse economy. While also bringing feedback to human users through the Metaverse infrastructure, the Metaverse engine pushes multi-sensory multimedia content to the Metaverse [170]. Through their avatars, users interact with the Metaverse by requesting services from one or more virtual worlds. The Metaverse's virtual worlds are enabled by the sensing and actuation infrastructure layer, where the ubiquitous IoT and sensor networks – installed in the physical world – gather gigantic data from the physical environment [171]. To mitigate the impact of such big data that can cause performance bottleneck to the 6G Metaverse, (goal-oriented) SemCom plays a paramount role by transmitting only semantically-relevant information that informs the effective execution of an action, while minimizing bandwidth consumption, power usage, and transmission delay.

Driven by the overarching paradigm shift toward semantic-aware and task-oriented communications, 6G communication trends such as *task-oriented and semantics-aware communication* [172], *task-oriented communication design* [88], *task-oriented integrated sensing, computation, and communication in edge AI inference* [173], *task-oriented explainable SemCom* [174], *cooperative goal-oriented SemCom* [175], *multi-user goal-oriented SemCom* [176], *neuro-symbolic AI for intent-based goal-oriented SemCom* [177], and *goal-oriented SemCom with AI tasks* [178] have emerged. Meanwhile, the ongoing developments in SemCom, goal-oriented SemCom, and 6G are mutually reinforcing [60]. This motivates the following discussion on why 6G is crucial for goal-oriented SemCom.

D. WHY 6G FOR GOAL-ORIENTED SEMCOM?

The ongoing developments of 6G are core enablers – and thus opportunities – for further development of goal-oriented SemCom [60]. Particularly, goal-oriented SemCom can considerably benefit from the emergence of *computing*

[5] These 6G use cases are also promising for M2M goal-oriented SemCom, which will be essential for the design and materialization of 6G like H2H goal-oriented SemCom and H2M goal-oriented SemCom.

force networks (CFN), *AI-native networks*, *ubiquitous connectivity*, and *trustworthiness-native networks* [60], which are henceforth discussed below.

1) COMPUTING FORCE NETWORK

The success of mobile edge computing (MEC) deployment [179], [180], [181], [182] has facilitated the convergence of communication and computing, which has inspired [60] the emerging 6G paradigms of *computation-oriented communications* (see [5]) and CFN. CFN is anticipated to be an information system unifying the functions of communication, computing, and caching toward attaining mutual awareness, autonomous collaboration, and unified orchestration and management (O&M) [60]. CFN's benefits are [60]:

- Real-time accurate computing force (CF) detection.
- Flexible and dynamic scheduling of services.
- Consistency of user experience.

To rack up the itemized benefits, CFN needs *CF measurement and modeling*, *CF-awareness-based CF routing*, and *in-network computing* [60].

2) AI-NATIVE NETWORKS

To enhance data security, privacy, and overall network performance/efficiency, the 6G network has been envisioned to facilitate the following fundamental paradigm shifts in communications and computing [60]:

- Departure from cloud AI to distributed AI [183].
- The migration of data processing from the network core to the network edge [52], [179], [184].
- From connection-oriented communication to task-oriented communication [183] and computation-oriented communication [5], [57].

To enable the itemized paradigm shifts, AI-native networks (see [183], [185], [186], [187]) are proposed, where the AI-native architecture either leverages AI techniques to optimize network performance (i.e., *AI4Net* [188]), or provides real-time AI services for a broad range of industries (i.e., *Net4AI* [188]), dubbed *AI-as-a-service (AIaaS)* [60]. To materialize such AI-native architecture, *O&M for network AI* and *network function architecture for network AI* need to be implemented [60].

3) UBIQUITOUS CONNECTIVITY

To surmount the challenges of enormous data transmission (e.g., in holographic videos and collaborative autonomous driving) that will impose a considerable burden on the already resource-limited wireless communication networks, 6G networks must therefore be aimed at enabling ubiquitous connectivity [102], [189]. The above-mentioned paradigm shifts facilitate ubiquitous connectivity, which will, in turn, promote the development of 6G wireless systems that are based on goal-oriented SemCom. Such goal-oriented SemCom-based 6G systems can be designed and optimized to materialize global ubiquitous connectivity along the maturation of the following 6G technology enablers [60]:

- *SAGINs [48], [94], [95]*: Despite their challenges noted in Sec. II-A, they enable not only global broadband and massive IoT communication but also permit enhanced location navigation and real-time Earth observation [60].
- *(Sub-)THz communication [4], [27], [190], [191]*: It can support ultra-broadband communications that provide further guarantees for ubiquitous connectivity [60], though it faces many multi-faceted challenges highlighted in Sec. II-A (see also [192]).
- *Ultra-massive MIMO [50], [193]*: Enabled by AI, RIS, emerging sensing technologies, and ultra-large aperture array, ultra-massive MIMO has the potential to surmount the high transmission loss in the THz frequency band and achieve – in a broader frequency range – a higher spectral and energy efficiency, while being applicable to many use cases [60]. Nevertheless, this technology is confronted by the following challenges: High cost, the burden of signal processing load, limited fronthaul capacity, high reference signal overhead, and inaccurate channel measurement and modeling [60].

4) TRUSTWORTHINESS-NATIVE NETWORK

Due to the growing privacy and security concerns, a *trustworthy communication network* has become a key requirement in 6G [60]. 6G also requires the evolution of the network architecture (from centralized control to distributed processing) along with a tighter convergence of AI and big data, leading to an inevitable and huge exchange of data [60]. This certainly puts forth new challenges for 6G security, privacy, and trust [60]. Accordingly, the trustworthiness-native network in 6G should take the following into consideration: Topology dynamic change, wide area network sharing mechanism, access control model, and technologies of isolation and exchange [60]. In addition, the 6G network will require a native security architecture with *autonomous defense capabilities*, considering the trends of convergence and integration in 6G [60].

Despite the many challenges and open problems – for the 6G security, privacy, and trust – presented by Table 7, the design of trustworthiness-native networks is going to be at the forefront of 6G research [5], [11], which will in turn facilitate the realization of goal-oriented SemCom through secure-by-design 6G networks. Apart from secure-by-design 6G networks, the realization of major goal-oriented SemCom use cases requires the design and optimization of a (possibly) autonomous and integrated 6G network. Therefore, the paradigms of *6G for goal-oriented SemCom* and *goal-oriented SemCom for 6G* necessitate a tighter integration of 6G and goal-oriented SemCom. In view of this purpose, we now move on to expound the state-of-the-art research landscape of goal-oriented SemCom.

III. STATE-OF-THE-ART RESEARCH LANDSCAPE OF GOAL-ORIENTED SEMCOM

When looking beyond conventional wireless connectivity, it is worth underscoring that communication is not an end in

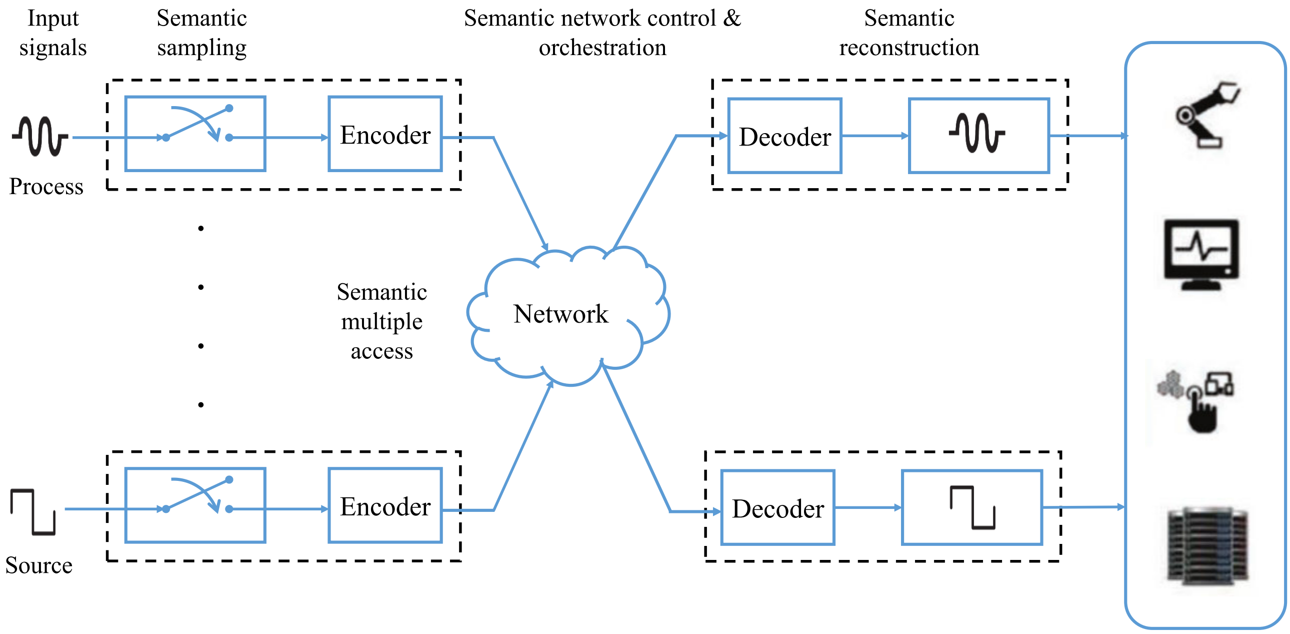


FIGURE 5. End-to-end goal-oriented SemCom model – redrawn from [83, Fig. 1].

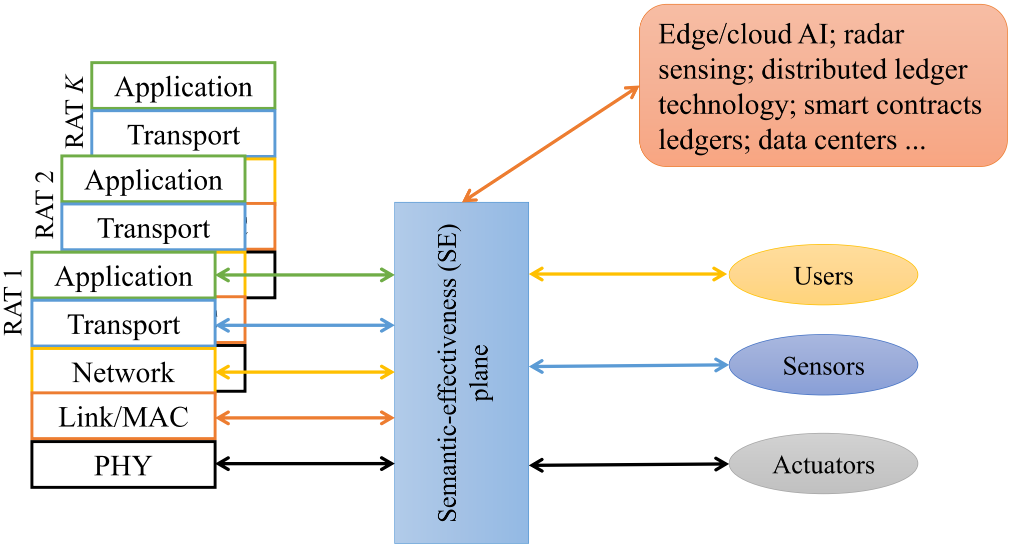

Edge/cloud AI; radar sensing; distributed ledger technology; smart contracts ledgers; data centers ...
RAT K
Application
Transport
RAT 2
Application
Transport
RAT 1
Application
Transport
Network
Link/MAC
PHY
Semantic-effectiveness (SE) plane
Users
Sensors
Actuators

FIGURE 6. A goal-oriented SemCom architecture with semantic-effectiveness filtering – redrawn from [194, Fig. 1(b)]. RAT: radio access technology.

itself, but a means to achieving definite goals [83]. The end-to-end goal-oriented SemCom model that is proposed by the authors of [83] and depicted in Fig. 5 is therefore crucial. This figure comprises the following four building blocks.

- *Multiple continuous or discrete signals (stochastic processes);* various (possibly correlated) signals illustrating time-varying real-world physical phenomena in space are observed by spatially distributed smart devices, which are empowered by heterogeneous sensing, computational, and learning/inference capabilities [83].
- *A shared communication medium:* a shared medium is used jointly by smart devices to send data samples – e.g., their observations, measurements, and updates – to one or more destinations, such as a fusion center (FC) or a control unit [83]. Their respective samples are generated using process-aware (non-uniform active) sampling in accordance with the communication characteristics, the semantic-aware applications' requirements, and source variability (in terms of changes, innovation rate, autocorrelation, and self-similarity) [83].
- *Preprocessing of source samples:* prior to being encoded and scheduled for transmission over noisy and delay-prone (error-prone) communication channels, source samples could be preprocessed [83]. This preprocessing may incorporate quantization, compression, and feature extraction, among other processes [83].

For goal-oriented SemCom, meanwhile, scheduling is performed per the semantic information's value and priority, which are extracted from the input data [83].

- *Signal reconstruction:* the input signals are eventually reconstructed from causally or non-causally received samples at their respective destinations to serve the purpose of an application such as collision avoidance, remote state estimation, control and actuation, situation awareness, and learning model training [83].

Apart from the aforementioned early works on goal-oriented SemCom, the authors of [194] proffer a goal-oriented SemCom architecture (see Fig. 6) with a *semantic-effectiveness plane* [74, Fig. 3] whose functionalities address both the semantic and effectiveness problems. When it comes to these problems, and as schematized in Fig. 6, the architecture proposed in [194, Fig. 1(b)] supports not only information extraction, but also direct control.

We now move on to outline taxonomy for the state-of-the-art goal-oriented SemCom research.

A. TAXONOMY FOR THE STATE-OF-THE-ART GOAL-ORIENTED SEMCOM RESEARCH

We have carried out extensive research on all aspects of goal-oriented SemCom, which was informed by our search methodology that employed *Google Scholar* with key words "Goal-oriented semantic communication'', "Goal-oriented communication'', "Task-oriented communication'', and "Task-oriented semantic communication''. This search has led to our taxonomy depicted in Fig. 7, which informs our entire survey and discussions on:

- State-of-the-art works of goal-oriented SemCom.
- Major trends of goal-oriented SemCom.
- Major use cases of goal-oriented SemCom.
- Mathematical frameworks of goal-oriented SemCom.
- Theories of goal-oriented SemCom.
- Challenges of goal-oriented SemCom.
- Future directions of goal-oriented SemCom.

In line with our taxonomy per Fig. 7, we now proceed to state-of-the-art vision and tutorial works on goal-oriented SemCom.

B. VISION AND TUTORIAL WORKS ON GOAL-ORIENTED SEMCOM

We highlight below vision and tutorial works on goal-oriented SemCom, beginning with vision works.

1) VISION WORKS ON GOAL-ORIENTED SEMCOM

The authors of [83] envision a communication paradigm shift that requires the goal-oriented unification of information generation, information transmission, and information reconstruction while taking into account multiple factors such as process dynamics, data correlation, signal sparsity, and semantic information attributes. The authors of [195] present a vision of a new paradigm shift that targets joint optimal information gathering, information dissemination, and decision-making policies in NCSs that incorporate the semantics of information based on the significance of the messages – but not necessarily the meaning of the messages, and possibly with a real-time constraint – w.r.t. the purpose of the data exchange. The authors of [59] present their vision of 6G wireless networks, wherein SemCom and goal-oriented SemCom are promising technologies that derive a crucial paradigm shift away from Shannon's information-theoretic framework. This paradigm shift underscores the fact that the success of task execution at a given destination (the effectiveness problem) is more of the essence than achieving error-free communication at the symbol level (the technical problem) [59].

To ensure the concrete representation and efficient processing of the semantic information, the authors of [159] introduce a formal graph-based semantic language and a goal filtering method for goal-oriented signal processing. Expanding upon this framework, the authors of [161] introduce a semantic information extraction framework wherein the extracted graph-based imperfect semantic signals can be improved for better fidelity and filtered for reduced semantic source noise. The authors of [196] put forward an architecture that makes it possible to learn the representation of semantic symbols for goal-oriented SemCom (effectiveness-level SemCom) and design objective functions, which would help train effective semantic encoders/decoders. The authors of [197] present the challenges and opportunities related to goal-oriented SemCom networks while advocating goal-oriented SemCom as an enabler of 6G use cases. The authors of [198] present a 6G vision based on task-oriented communication, whose effectiveness is demonstrated – by the same authors – for federated learning, edge inference, and SemCom.

We now proceed to highlight the existing tutorial works on goal-oriented SemCom.

2) TUTORIAL WORKS ON GOAL-ORIENTED SEMCOM

The authors of [60] provide a partial review of the fundamentals, applications, and challenges of goal-oriented SemCom. The authors of [78] offer a tutorial – for communication theorists and practitioners – that provides an introduction to state-of-the-art tools and advancements in goal-oriented SemCom. The authors of [85] offer a partial overview of recent research developments in goal-oriented SemCom while focusing on goal-oriented data compression for IoT applications. The authors of [86] review goal-oriented SemCom and semantic transformations.

Apart from the aforementioned vision and tutorial works on goal-oriented SemCom, the rapidly evolving state-of-the-art works on goal-oriented SemCom investigate numerous goal-oriented SemCom techniques, trends, and use cases such as task-oriented communication with digital modulation [87]; goal-oriented SemCom with AI tasks [178]; intent-based goal-oriented SemCom [177], [199]; multi-user goal-oriented SemCom [176]; and cooperative SemCom [175].

We now move on to state-of-the-art algorithmic developments in goal-oriented SemCom.

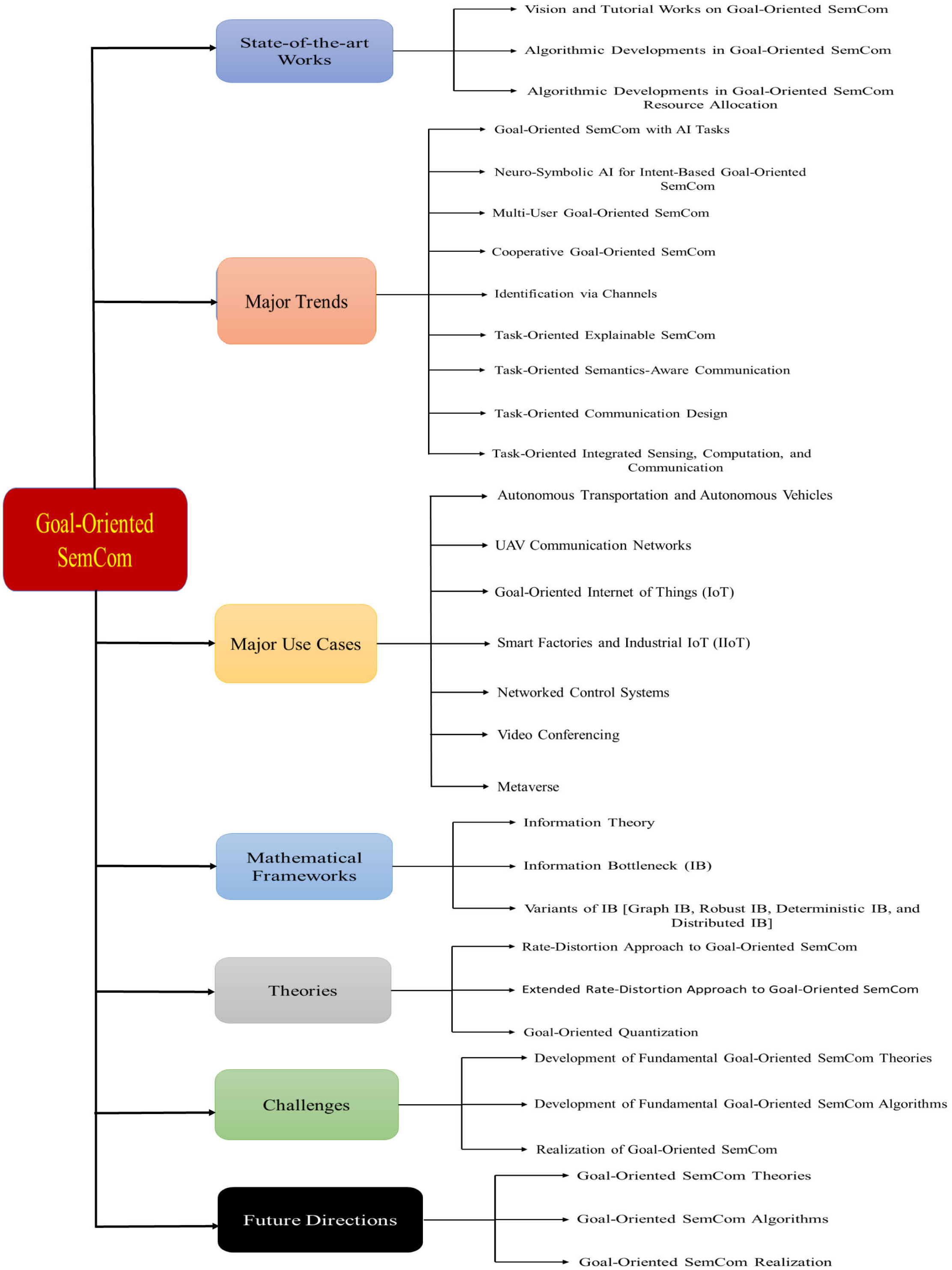

Goal-Oriented SemCom
State-of-the-art Works
Vision and Tutorial Works on Goal-Oriented SemCom
Algorithmic Developments in Goal-Oriented SemCom
Algorithmic Developments in Goal-Oriented SemCom Resource Allocation
Major Trends
Goal-Oriented SemCom with AI Tasks
Neuro-Symbolic AI for Intent-Based Goal-Oriented SemCom
Multi-User Goal-Oriented SemCom
Cooperative Goal-Oriented SemCom
Identification via Channels
Task-Oriented Explainable SemCom
Task-Oriented Semantics-Aware Communication
Task-Oriented Communication Design
Task-Oriented Integrated Sensing, Computation, and Communication
Major Use Cases
Autonomous Transportation and Autonomous Vehicles
UAV Communication Networks
Goal-Oriented Internet of Things (IoT)
Smart Factories and Industrial IoT (IIoT)
Networked Control Systems
Video Conferencing
Metaverse
Mathematical Frameworks
Information Theory
Information Bottleneck (IB)
Variants of IB [Graph IB, Robust IB, Deterministic IB, and Distributed IB]
Theories
Rate-Distortion Approach to Goal-Oriented SemCom
Extended Rate-Distortion Approach to Goal-Oriented SemCom
Goal-Oriented Quantization
Challenges
Development of Fundamental Goal-Oriented SemCom Theories
Development of Fundamental Goal-Oriented SemCom Algorithms
Realization of Goal-Oriented SemCom
Future Directions
Goal-Oriented SemCom Theories
Goal-Oriented SemCom Algorithms
Goal-Oriented SemCom Realization

FIGURE 7. Taxonomy for the state-of-the-art goal-oriented SemCom research.

C. ALGORITHMIC DEVELOPMENTS IN GOAL-ORIENTED SEMCOM

We hereunder detail contemporary algorithms for single-user/single-task goal-oriented SemCom and multi-user/multi-task goal-oriented SemCom, starting with the former.

1) ALGORITHMS FOR SINGLE-USER/SINGLE-TASK GOAL-ORIENTED SEMCOM

The authors of [200] aim to devise a joint sampling and communication scheme over a wireless multiple access channel to compute the empirical probability measure of a quantity of interest at the destination and put forward a goal-oriented SemCom strategy that encompasses both (*i*) semantic-aware active sampling for goal-oriented signal reconstruction (at a FC) and (*ii*) a transmission scheme to access the shared communication medium. The authors of [201], on the other hand, propose a semantic information-aware policy for a MIMO-OFDM system – aimed at image classification – that is employed to transmit images to multiple users. In this goal-oriented SemCom system that is made up of a convolutional neural network (CNN)-based transmitter and a CNN-based receiver, a graph neural network that is fed modulated symbols is deployed to learn a precoding policy [201]. This policy is demonstrated to outperform regularized zero-forcing precoding and zero-forcing precoding in minimizing the bandwidth consumed by required data to attain an expected level of classification accuracy [201].

The authors of [178] underscore the premise that SemCom must take AI tasks into account and put forward a goal-oriented SemCom paradigm dubbed *SemCom paradigm with AI tasks* (SC-AIT), which is schematized in Fig. 8. Inspired by this goal-oriented SemCom systems (among others), the authors of [202] investigate a goal-oriented SemCom scheme for image classification task offloading in aerial systems in addition to proffering a joint SE-compression model. Their system is demonstrated to deliver an optimal SE under various channel states while taking into consideration the system's optimization objective that comprises the uplink transmission latency and the classification accuracy of the back-end target model [202]. Moreover, the authors of [203] proffer a curriculum learning-based SemCom framework for goal-oriented task execution. Building on this work, the authors of [204] introduce a goal-oriented SemCom model that incorporates a speaker and a listener who wish to jointly execute a set of tasks for task execution in a dynamic environment with the objective of jointly optimizing task execution time, transmission cost, inference cost, and resource efficiency. To solve this optimization problem, the authors of [204] provide a reinforcement learning (RL)-based bottom-up curriculum learning framework shown to outperform traditional RL in terms of convergence time, task execution cost and time, reliability, and belief efficiency [204].

In view of emerging 6G applications such as XR online role-playing game, the authors of [205] proffer a MEC structure for goal-oriented multimodal SemCom, wherein the proposed structure deploys a bidirectional caching task model (a realistic model for emerging AI-enabled applications). More specifically, the authors of [205] put forward an offloading scheme with cache enhancement to minimize a system's computation cost by formulating the cache-computational resource coordination problem as a mixed integer non-linear programming problem. As a result, they develop the content popularity-based deep Q-network (DQN) caching algorithm (CP-DQN) to make quasi-optimal caching decisions and the cache-computing coordination algorithm (CCCA) to achieve a tradeoff between using computing resources and caching [205]. The CP-DQN and CCCA algorithms are shown to perform optimally w.r.t. cache hit rate, cache reward, and system cost reduction [205].

Contrastingly, many current works are geared toward designing advanced algorithms for high-performance goal-oriented SemCom [206]. Nonetheless, energy-hungry and efficiency-limited image retrieval and semantic encoding without considering user personality are major challenges for unmanned aerial vehicle (UAV) image-sensing-driven goal-oriented SemCom scenarios [206]. To overcome these challenges, the authors of [206] devise an energy-efficient goal-oriented SemCom framework that uses a triple-based scene graph for image information. Meanwhile, the authors of [206] develop a personalized attention-based mechanism to achieve the differential weight encoding of triplets for crucial information following user preferences and ensure personalized SemCom. This scheme's ability to achieve personalized SemCom is corroborated by numerical results [206].

The authors of [207] leverage the *information bottleneck* (IB) [208] framework (see Section V) to formalize a rate-distortion tradeoff between the encoded feature's informativeness and the inference performance and design a goal-oriented SemCom system. They incorporate variational approximation – named variational IB (VIB) – in their system to build a tractable upper bound w.r.t. IB optimization, which is computationally prohibitive for high-dimensional data. Meanwhile, their system is shown to achieve a better rate-distortion tradeoff than baseline methods while considerably reducing feature transmission latency in dynamic channel conditions [207]. Building on the work in [207], the authors of [209] put forward a goal-oriented SemCom strategy for multi-device cooperative edge inference wherein a group of edge devices transmit task-relevant features to an edge server for aggregation and processing by leveraging the IB framework [208] and the distributed IB (DIB) framework [210]. The IB framework and the DIB framework are exploited in [209] for feature extraction and distributed feature encoding, respectively. This IB- and DB-based goal-oriented SemCom technique is shown to significantly reduce communication overhead in comparison with conventional data-oriented communication and, in turn, enable low-latency cooperative edge inference [209]. Building on the work in [209] and in [207], the authors of [211] study goal-oriented SemCom for edge video analytics by exploiting the deterministic IB framework [212] for feature extraction and the temporal entropy model for encoding. This goal-oriented SemCom scheme outperforms conventional data-oriented communication strategies in terms of its rate-performance tradeoff [211].

Corresponding to the effectiveness level of Weaver's three levels of communication (see Fig. 2), the authors of [213] investigate a multi-agent partially observable Markov decision process (MA-POMDP), wherein agents not only interact with the environment but also communicate with each other over a noisy communication channel. In light of this multi-agent RL (MARL) framework, the authors of [213] demonstrate that the joint policy that is learned by all the agents is far better than the one that is obtained by treating the communication and principal MARL problems separately.

To minimize the amount of semantic information needing to be transmitted for a given task, many works on goal-oriented SemCom aim to transmit only task-relevant information without introducing any redundancy. Nevertheless, doing so with a joint source-channel coding (JSCC)-based design causes robustness issues in learning due to channel variation and JSCC, while mapping the source data directly to continuous channel input symbols poses compatibility issues with existing digital communication systems [87]. To address these challenges while examining the inherent tradeoff between the informativeness of the encoded representations and the robustness of the received representations to information distortion, the authors of [87] devise a goal-oriented SemCom system with digital modulation that is dubbed *discrete task-oriented JSCC* (DT-JSCC). In DT-JSCC, the transmitter encodes the extracted input features into a discrete representation and transmits it to the receiver using digital modulation [87]. As for the DT-JSCC scheme's improved robustness to channel variation, the authors of [87] develop an IB-based encoding framework named *robust IB* (RIB) and derive a tractable variational upper bound for the RIB objective function using variational approximation [87]. Consequently, DT-JSCC is shown to be robust against channel variation with better inference performance than low-latency baseline methods [87].

The authors of [214] leverage the significance and effectiveness of messages to devise new goal-oriented sampling and communication policies as a means of generating and transmitting only the "most informative samples" for real-time tracking in autonomous systems. For these systems, the results reported by the authors of [214] demonstrate that semantics-empowered policies considerably reduce real-time reconstruction error, the cost of actuation error, and the amount of ineffective updates [214, Sec. 5].

We now continue to state-of-the-art algorithms for multi-user/multi-task goal-oriented SemCom.

2) ALGORITHMS FOR MULTI-USER/MULTI-TASK GOAL-ORIENTED SEMCOM

In single-user goal/task-oriented SemCom, either the trained model has to be updated once the task is altered or several trained models need to be stored to serve different tasks [215]. To overcome this limitation, the authors of [215] develop a unified DL-enabled SemCom system named *U-DeepSC*. U-DeepSC is a unified end-to-end framework that is designed to serve various tasks with multiple modalities [215], [216]. Moreover, the authors of [215] devise a multi-exit architecture in U-DeepSC to provide early-exit results for relatively simple tasks and design a unified codebook for feature representation to serve different tasks with reduced transmission overhead.

Aiming to exploit multimodal data from multiple users, the authors of [217] propose a multi-user task-oriented SemCom system for visual question answering (VQA), named *MU-DeepSC*, to exploit multimodal data from multiple users. MU-DeepSC is a DL-enabled goal-oriented SemCom system whose transceiver is designed and optimized to jointly capture features from the correlated multimodal data of multiple users [217]. Consequently, MU-DeepSC is demonstrated to be more robust to channel variation than traditional communication systems, especially in low signal-to-noise ratio (SNR) regimes [217]. Building on the work in [217], the authors of [176] design and implement multi-user task-oriented SemCom systems for the transmission of both data with one modality and data with multiple modalities. The authors consider image retrieval / machine translation for their single-modal task and VQA for their multimodal task [176]. The authors of [176] develop three Transformer-based transceivers for their systems, which are dubbed *DeepSC-IR*, *DeepSC-MT*, and *DeepSC-VQA*, that share the same transmitter structure but have different receiver structures [176]. When these transceivers are trained jointly in an end-to-end manner using the training algorithms in [176], they are corroborated to outperform traditional transceivers, especially in low SNR regimes [176].

Apart from the above-detailed algorithmic developments in goal-oriented SemCom, useful algorithmic developments have also been made in goal-oriented SemCom resource allocation, which we discuss below.

D. ALGORITHMIC DEVELOPMENTS IN GOAL-ORIENTED SEMCOM RESOURCE ALLOCATION

The authors of [218] consider a multi-user goal-oriented SemCom system at the wireless edge and exploit the Lyapunov optimization framework to devise a joint computation and transmission management strategy for their overall system. More specifically, the authors of [218] develop a multi-user minimum energy resource allocation strategy that ensures energy-efficient optimal resource allocation for edge devices and edge server. This resource allocation strategy's simulation results demonstrate there is an edge-ML trade-off between energy, latency, and accuracy [218]. Extending the work in [218], the authors of [219] investigate the trade-offs between energy, latency, and accuracy in a goal-oriented SemCom-enabled edge learning system. More specifically, they develop two resource optimization strategies (that also exploit the Lyapunov stochastic optimization framework) to jointly optimize the communication parameters and the computation resources while aiming for an optimal trade-off between energy, latency, and accuracy for the edge learning task [219]. These proposed strategies are corroborated, by simulations, to provide adaptation capabilities and to be effective for edge learning with goal-oriented SemCom [219].

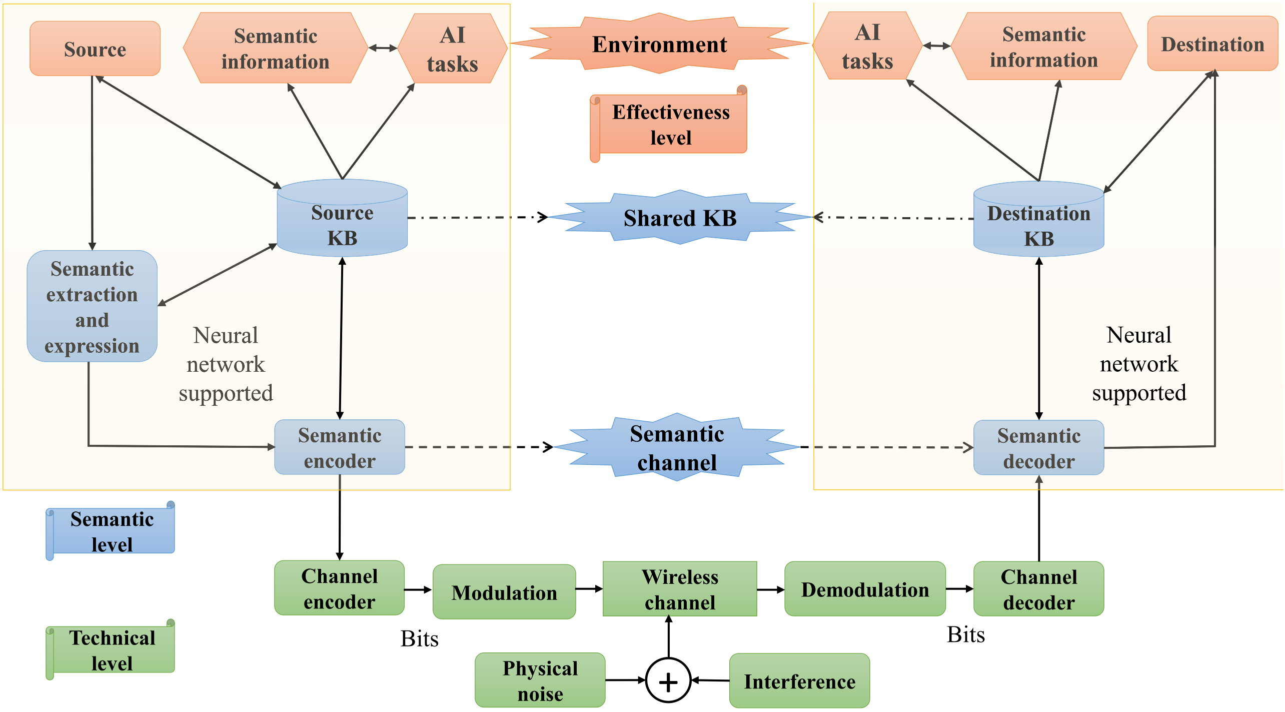


FIGURE 8. Goal-oriented SemCom with AI tasks – redrawn from [178, Fig. 2].

When it comes to personalized saliency-based goal-oriented SemCom in UAV image sensing scenarios, the authors of [206] investigate SemCom personalization and its corresponding optimal resource allocation. For the former, the authors theoretically analyze the effects of wireless fading channels on SemCom, and for the latter, they put forward a game-based model for multi-user resource allocation (to efficiently utilize UAV resources). This framework is confirmed to improve UAV resource utilization [206].

The authors of [220] propose a multi-user goal-oriented SemCom framework that aims to enable users to effectively extract, compress, and transmit the semantics of input data to the edge server. The edge server then executes the intelligence task based on the received semantics and delivers results to users [220]. Meanwhile, the authors of [220] also propose a new approach dubbed *adaptable semantic compression* (ASC) to compress the extracted semantics based on semantic importance, which helps to reduce the communication burden. However, ASC faces the following problem in a multi-user setting: a higher compression ratio requires fewer channel resources but causes considerable semantic distortion, while a lower compression ratio calls on more channel resources and hence results in transmission failure due to the delay constraint (especially in delay-intolerant systems) [220]. In light of this problem, the authors of [220] formulate a resource allocation and compression ratio optimization problem that aims to maximize the *success probability of tasks*[6] under bandwidth and power constraints. In addressing this non-convex problem, the authors of [220] develop two algorithms that achieve greater task performance gains than the baseline algorithms do while significantly reducing the volume of data transmitted [220, Sec. VI].

The authors of [169] disseminate a goal-oriented system design for the IoT, considering both architectural and resource optimization aspects, after investigating the role of sensing, communication, and computing resources in attaining the tradeoff between goal cost and effectiveness. To tackle this goal-oriented optimization problem, the authors of [169] leverage Lyapunov stochastic optimization to learn optimal policies that can incorporate models describing the system operation, even under unknown statistics of channel state or request arrival rates. In light of the work in [169], the authors of [222] analyze and optimize a wireless network coexistence scenario between a goal-oriented communication system and a legacy data-oriented communication system, while fully sharing the same spectrum resources.

We now continue to major state-of-the-art trends and use cases of goal-oriented SemCom.

IV. MAJOR STATE-OF-THE-ART TRENDS AND USE CASES OF GOAL-ORIENTED SEMCOM

In this section, we present the major state-of-the-art trends and use cases related to goal-oriented SemCom, beginning with the major trends.

A. MAJOR TRENDS OF GOAL-ORIENTED SEMCOM

We discuss the following major trends of goal-oriented SemCom: *goal-oriented SemCom with AI tasks* [178],

[6] The success probability of tasks is defined to quantify the performance of goal-oriented SemCom systems and is given by [221, eq. (11)].

neuro-symbolic AI for intent-based goal-oriented SemCom [177], *multi-user goal-oriented SemCom* [176], *cooperative goal-oriented SemCom* [175], *identification via channels* [223], [224], [225], *task-oriented explainable SemCom* [174], *task-oriented and semantics-aware (TOSA) communication* [172], *task-oriented communication design (TOCD)* [88], and *task-oriented integrated sensing, computation, and communication (ISCC) in edge AI inference* [173]. We start our discussion with goal-oriented SemCom with AI tasks.

1) GOAL-ORIENTED SEMCOM WITH AI TASKS

The authors of [178] were the first to assert that semantic information is closely related to the target AI task. This assertion is indeed reasonable when one considers the detection of a dog from a transmitted image that comprises both a dog and a cat (see [178, Fig. 1]), since the information related to the cat is no longer relevant. For this goal-oriented SemCom scenario, the authors of [178] put forward a goal-oriented SemCom system dubbed goal-oriented SemCom with AI tasks, which is shown in Fig. 8. Fig. 8 shows the technical and semantic levels – per Weaver's vision (shown in Fig. 2) – and a newly proposed effectiveness level. In their effectiveness level design, the authors propose to minimize the redundancy in the semantic information based on the contribution of raw information to the successful execution of AI tasks by discarding the information that is irrelevant to the success of AI tasks. This process can be conducted per the knowledge stored in the source KB that can be designed to account for the relationships between the AI tasks and the semantic information [178].

Once encoded using a semantic encoder, the semantic information is then is then channel-encoded and modulated prior to its transmission over a wireless channel. The received semantic information, which may be contaminated by physical noise and interference is then demodulated and channel-decoded, as seen in Fig. 8. This information is fed to a semantic-level receiver (as shown in Fig. 8), whose semantic decoder is employed to recover the transmitted semantic information in accordance with the destination KB. The destination KB can *synchronize* its knowledge elements with those of the source KB through a shared KB that can be stored in an authoritative third party or a virtual KB [178].

We now proceed to neuro-symbolic AI for intent-based goal-oriented SemCom [177].

2) NEURO-SYMBOLIC AI FOR INTENT-BASED GOAL-ORIENTED SEMCOM

In contrast with the state-of-the-art works on goal-oriented SemCom that characteristically lack data explainability, the work in [177] leverages *neuro-symbolic AI* (NeSy AI) [144], [147] and generative flow networks (GFlowNets) [226] to introduce a goal-oriented SemCom model named *NeSy AI* that aims to bring intelligence to the end nodes and is depicted in Fig. 9. As is shown in Fig. 9, NeSy AI's transmitter comprises an attribute extraction module, a state description module, and an encoder. When it comes to the latter two components, the state description module is learnable using neural AI and grounded in real logic according to the semantic language rules that are embedded in symbolic AI, and the encoder is realizable using neural AI and translates the states to (optimal) physical messages [177]. The receiver, on the other hand, is made up of a decoder and a semantic state extraction module, as shown in Fig. 9. As can be seen in Fig. 9, the decoder (which is designed using neural AI) transforms the received message into an estimated state description that is fed to the SE module (which is also designed using neural AI) that effectively recovers the transmitted semantic states in accordance with the reference semantic language rules (which are realizable using symbolic AI) [177].

In NeSy AI, the symbolic part is elaborated by the KB, and the reasoning – from learning the probabilistic structure that generates the data – is enabled by the GFlowNet [226]. The authors of [177], thus, formulate an optimization problem for causal structure learning – from the data and the optimal encoder/decoder functions – whose simulation results indicate it needs to transmit considerably fewer bits than a conventional communication system to reliably convey the same meaning [177]. Building on the work in [177] and NeSy AI, the authors of [199] introduce a goal-oriented SemCom framework named *emergent semantic communication* (ESC). ESC is made up of a signaling game for emergent language design and a NeSy AI approach for causal reasoning [199]. To design an emergent language that is compositional and semantic-aware, the authors of [199] solve the signaling game – using alternating maximization between the transmit and receive nodes' utilities – and characterize the generalized Nash equilibrium. The authors also deploy GFlowNet [226] to induce causal reasoning at the nodes and prove analytically that ESC systems can encode data with minimal bits in comparison with a classical system that does not employ causal reasoning [199].

We now move on to our discussion of multi-user goal-oriented SemCom systems [176].

3) MULTI-USER GOAL-ORIENTED SEMCOM

The authors of [176] devise a multi-user goal-oriented SemCom system, which is depicted in Fig. 10, to extend the benefits of single-user, single-modal goal-oriented SemCom to multiple users. Their proposed system is a multi-user MIMO system that is made up of a receiver equipped with M antennas and K single-antenna transmitters [176]. Each of the transmitters consists of a DL-based semantic encoder and a JSC encoder (both of which are learnable in an end-to-end fashion), and accepts image, text, video, or speech signals as input [176]. The receiver, on the other hand, can either be a *single-modal multi-user semantic receiver* and enable single-modal multi-user data transmission or be a *multimodal multi-user semantic receiver* and enable multimodal multi-user data transmission, as can be seen in Fig. 10 [176].

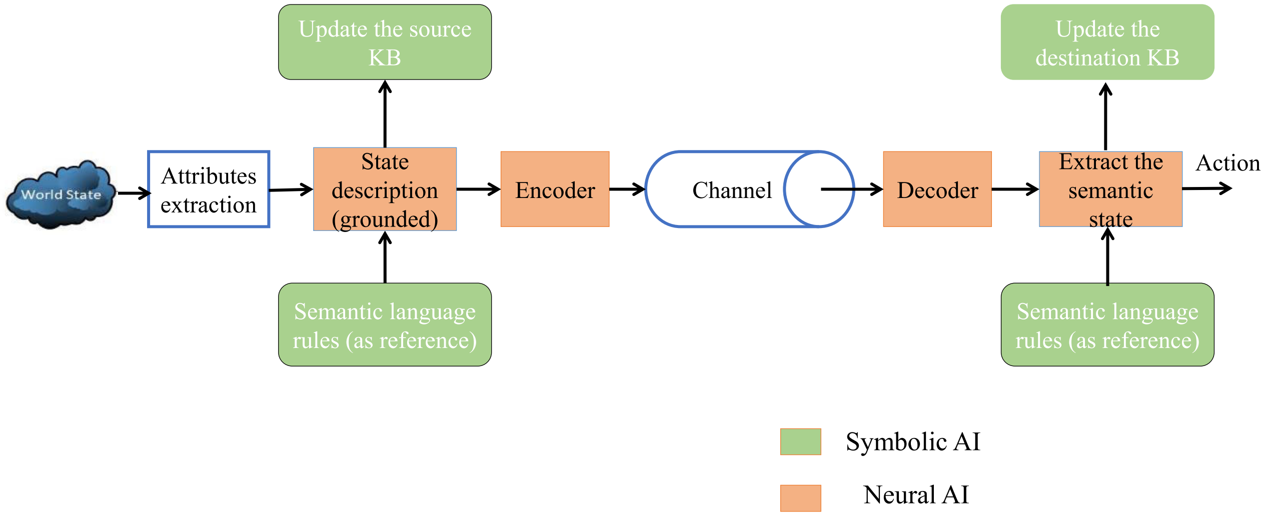

Update the source KB
Update the destination KB
World State
Attributes extraction
State description (grounded)
Encoder
Channel
Decoder
Extract the semantic state
Action
Semantic language rules (as reference)
Semantic language rules (as reference)
🟩 Symbolic AI
🟧 Neural AI

FIGURE 9. Intent-based goal-oriented SemCom (NeSy AI) – redrawn from [177, Fig. 1].

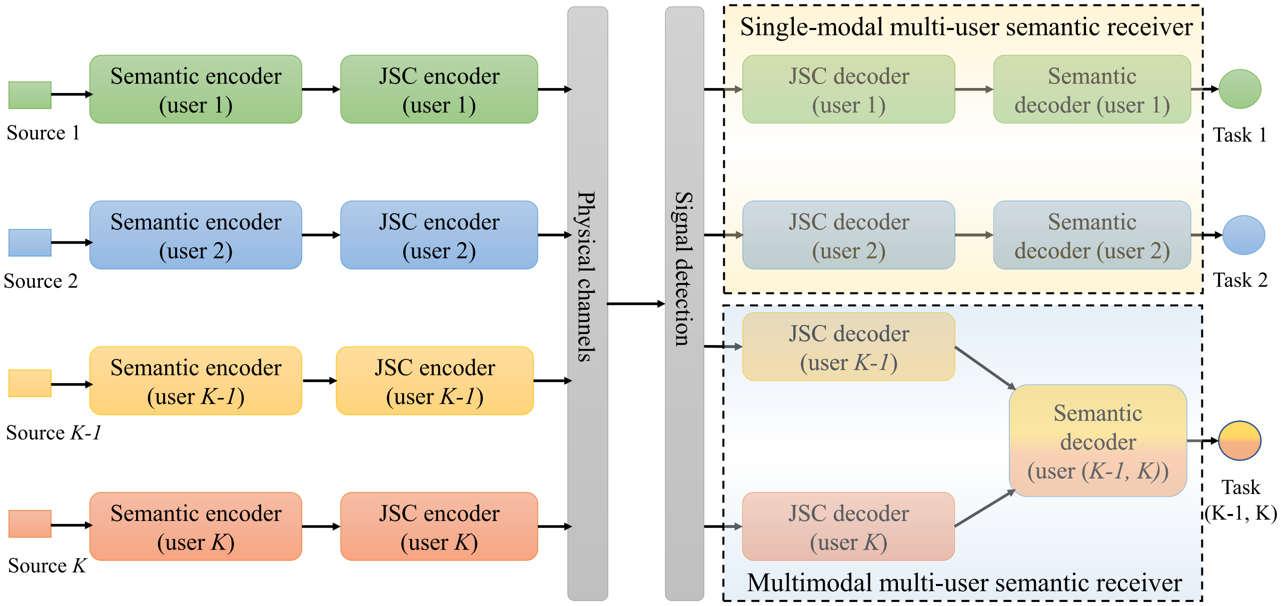

Source 1
Semantic encoder (user 1)
JSC encoder (user 1)
Source 2
Semantic encoder (user 2)
JSC encoder (user 2)
Source K-1
Semantic encoder (user K-1)
JSC encoder (user K-1)
Source K
Semantic encoder (user K)
JSC encoder (user K)
Physical channels
Signal detection
Single-modal multi-user semantic receiver
JSC decoder (user 1)
Semantic decoder (user 1)
Task 1
JSC decoder (user 2)
Semantic decoder (user 2)
Task 2
JSC decoder (user K-1)
Semantic decoder (user (K-1, K))
Task (K-1, K)
JSC decoder (user K)
Multimodal multi-user semantic receiver

FIGURE 10. Multi-user goal-oriented SemCom systems – redrawn from [176, Fig. 1]. JSC: joint source-channel.

Single-modal multi-user transmission means that each user independently transmits its extracted semantic information to carry out its task [176]. Multimodal multi-user transmission, on the other hand, means that the data from different users are semantically complementary [176]. Each of these goal-oriented SemCom systems relies on the linear minimum-mean-squared error (L-MMSE) detector to recover signals with estimated channel state information (CSI) [176]. Each user's JSC decoder is designed/trained to decompress the received semantic information following L-MMSE detection while mitigating the effects of channel distortion and inter-user interference [176]. When the JSC decoder is used in sequence with the semantic decoder to form a single-modal multi-user semantic receiver, as schematized in Fig. 10, each user's semantic information is exploited to perform different tasks independently [176]. This single-modal multi-user goal-oriented SemCom system can be used for the joint performance of an image retrieval task and a machine translation task, as in [176, Fig. 2]. Moreover, as is shown in Fig. 10, the final task that corresponds to a multimodal semantic receiver is completed by merging the different users' semantic information [176]. This multi-user goal-oriented SemCom system is useful for realizing a multimodal multi-user goal-oriented SemCom system with a *DeepSC-VQA* transceiver, as shown in [176, Fig. 3].

The authors of [175] build on the multimodal multi-user goal-oriented SemCom system that is depicted in Fig. 10 and put forward a goal-oriented SemCom system

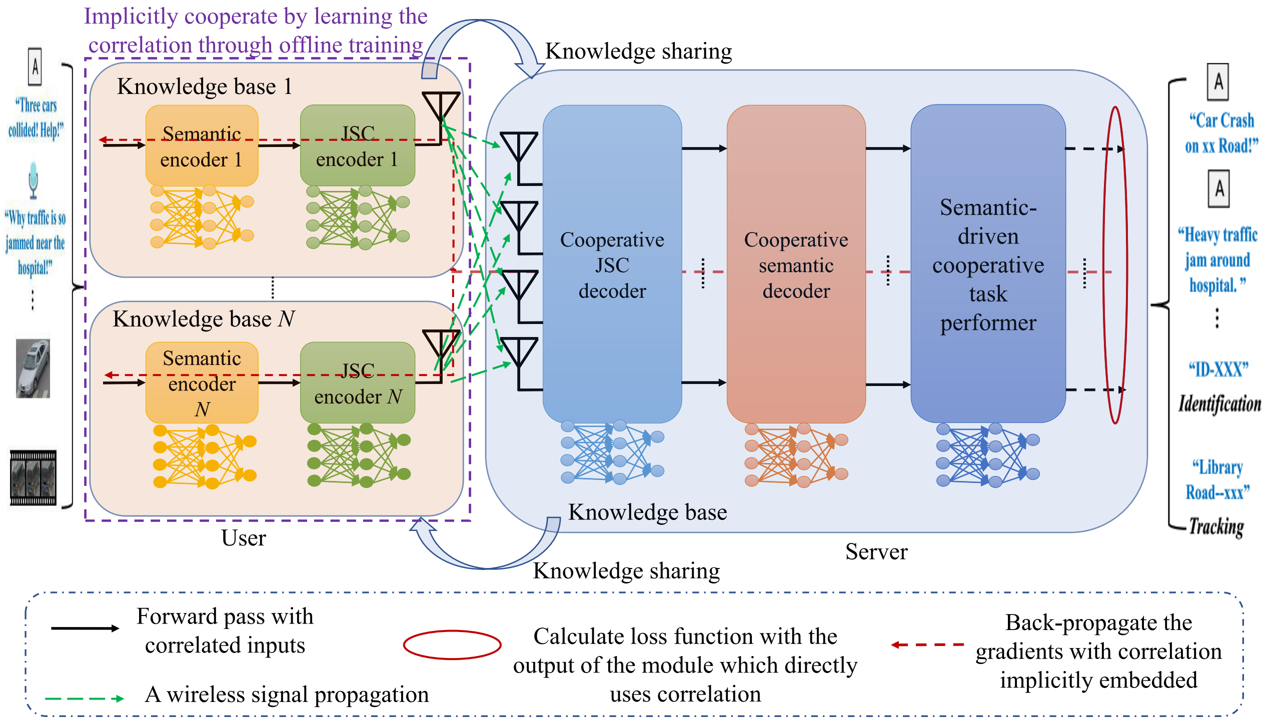

Implicitly cooperate by learning the correlation through offline training
Knowledge sharing
"Three cars collided! Help!"
"Why traffic is so jammed near the hospital!"
Knowledge base 1
Semantic encoder 1
JSC encoder 1
Knowledge base N
Semantic encoder N
JSC encoder N
User
Cooperative JSC decoder
Cooperative semantic decoder
Semantic-driven cooperative task performer
Knowledge base
Server
Knowledge sharing
"Car Crash on xx Road!"
"Heavy traffic jam around hospital."
"ID-XXX"
Identification
"Library Road--xxx"
Tracking
Forward pass with correlated inputs
A wireless signal propagation
Calculate loss function with the output of the module which directly uses correlation
Back-propagate the gradients with correlation implicitly embedded

FIGURE 11. An architecture for a general cooperative goal-oriented SemCom – redrawn from [175, Fig. 2]. JSC: joint source and channel.

named cooperative goal-oriented SemCom, which we discuss below.

4) COOPERATIVE GOAL-ORIENTED SEMCOM

It is proposed for the internet of vehicles (IoV) applications such as pedestrian detection, traffic analysis, and vehicle tracking [175]. A general cooperative goal-oriented SemCom architecture is shown in Fig. 11. As can be seen in Fig. 11, cooperative goal-oriented SemCom comprises a semantic encoder and a cooperative semantic decoder, a JSC encoder and a cooperative JSC decoder, and a semantic-driven cooperative task performer. Interestingly, the correlation among users is *pre-learned* and *embedded* in the cooperative goal-oriented SemCom architecture, including the encoders at the transmitters and the cooperative modules at the receiver [175]. The different modules' functions are itemized below.

- *Semantic encoder:* it is designed to extract semantic information from the source data with a focus on meaning and goal-relevance [175].
- *Cooperative semantic decoder:* it recovers the source data according to the specific goals set while leveraging semantic-level correlation among users [175].
- *JSC encoder:* it is applied to encode the extracted semantic information (the output of the semantic encoder) as channel input symbols [175].
- *Cooperative JSC decoder:* it is realized/trained to jointly recover the transmitted semantic information of multiple users, as depicted in Fig. 11.
- *Semantic-driven cooperative task performer:* it is used to achieve specific tasks/actions while adapting its structure to a specific task based on the semantic information recovered from multiple inputting users [175]. It also leverages semantic-level correlation and distinctive user attributes while cooperatively performing a task by combining information from different users [175].

The cooperative goal-oriented SemCom scheme shown in Fig. 11 requires that knowledge be shared between the transmitters and a receiver and that each user have a background KB [175]. The KBs are presumed to be shared between users and the server by jointly training the whole deep neural network (DNN) offline with a common dataset [175].

We now continue with identification via channels.

5) IDENTIFICATION VIA CHANNELS

Proposed and advocated first by Ahlswede in his seminal information theory works [223], [224], [225], *identification via channels* (ID) concerns the design of an identification receiver interested only in a particular message, as seen in Fig. 12 [227]. Contrary to a traditional receiver that is always interested in any message, as seen in Fig. 12, an identification receiver is only interested in message *j* – its chosen message, as viewed in Fig. 12, while rejecting all other different messages [227]. In this respect, an identification receiver does not attempt to figure out which message the sender tries to convey to it, but rather if the ID messages of the sender and the receiver are identical or not [228]. This is the goal of ID, making ID a goal-oriented communication paradigm [229].

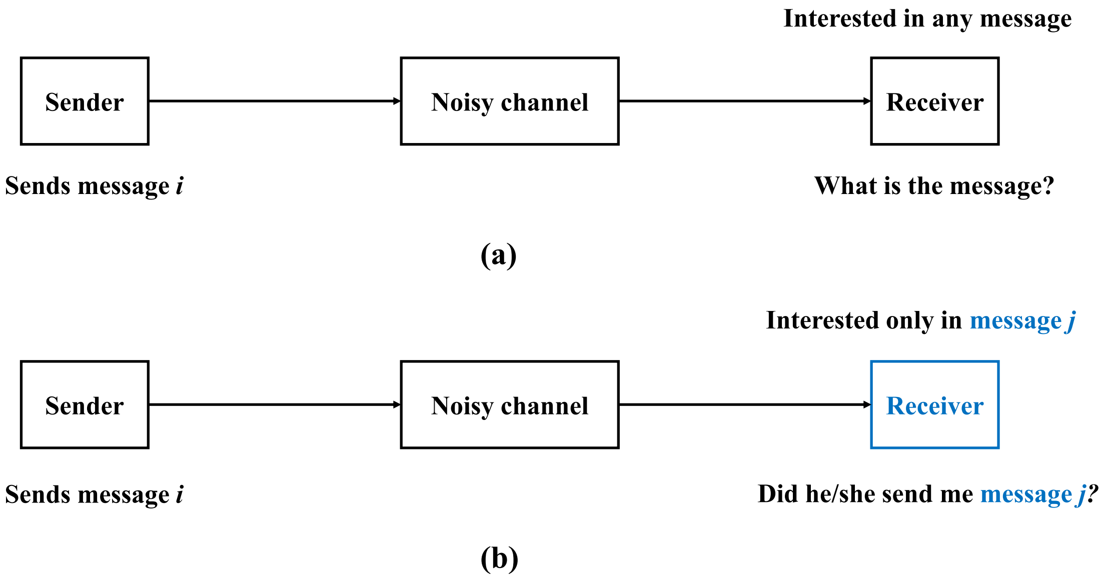


FIGURE 12. Schematics of a traditional and an identification receiver – redrawn from [227, Figs. 1 and 2]: (a) a traditional receiver interested in any message; (b) an identification receiver interested only in a particular message.

Unlike message transmission that is agnostic to the communication goal, ID encompasses knowledge of the sink's goal, i.e., the ID goal, and can attain exponential gains over message transmission [228]. Compared with the single-exponential scaling of Shannon's message transmission, ID makes the number of identifiable messages scale double exponentially with the block length and code rate [228]. Consequently, ID has an immense potential for distributed identity verification, which is indispensable in distributed databases [230], [231]; multi-access MEC [180], [181], [182]; and digital twins [162], [163], [164], [165]. By adding some pre-processing and post-processing at the sender and the receiver, respectively, identification codes can be constructed around transmission codes [227, Fig. 5]. In designing an identification receiver, two types of errors have to be independently dealt with: *missed identification* (an error of the first kind) that accounts for a regular transmission error and *false identification* (an error of the second kind) that is concerned with a false identification, where an identification receiver interested only in message *j* identifies a sender's transmitted message *i* as message *j* [227]. This is a *false positive error (FP-error)*, which occurs when the binary hypothesis test at the destination gives out a matching identification, though the messages at the source and destination are different [229]. Apart from the mentioned ID performance assessment metrics, the authors of [229] recently introduced a new ID performance metric named *the expected FP-error probability*.

We now move on to task-oriented explainable SemCom [174].

6) TASK-ORIENTED EXPLAINABLE SEMCOM

The authors of [174] proposed a new goal-oriented SemCom framework named task-oriented explainable SemCom that introduces a semantic-level transmission on top of a bit-level transmission, as schematized in Fig. 13. Per Fig. 13, the transmitter consists of a cascaded semantic encoder and feature selection modules that are connected to a conventional digital transmitter, made of a quantizer, a source encoder, and a channel decoder. As also seen in Fig. 13, the received signal is first processed by a traditional digital receiver – made of a channel decoder and source decoder – whose output is fed to a cascade of feature completion and semantic decoder modules. Accordingly, the framework of Fig. 13 is interoperable with the existing communication standards, protocols, and products [174].

The semantic encoder extracts the semantic features – w.r.t. the sender KB – of the source message X, and produces the disentangled features Z, while targeting to minimize the ambiguity incurred by the quantization error and channel noise [174]. The output of this module is fed to the feature selection module that aims to select – also w.r.t. the sender KB – only the task-relevant features for transmission while attempting to reduce the transmission load [174]. At the receiver, the feature completion module employs the source decoder's output and the receiver KB so as to compute a given task's target function and produce the estimated disentangled features $\hat{Z}$ [174]. Deploying $\hat{Z}$ and the receiver KB, the semantic decoder recovers the transmitted source data at the receiver. Moreover, the authors of [174] proposed the *rate-distortion-perception function* and *semantic channel capacity* to quantify the semantic information compression and transmission, respectively, and derived upper and lower bounds on the semantic channel capacity.

We now continue with TOSA communication [172].

7) TOSA COMMUNICATION

Proposed by the authors of [172], Fig. 14 shows a TOSA framework in which TOSA information is transmitted through a semantic network access to a receiver equipped

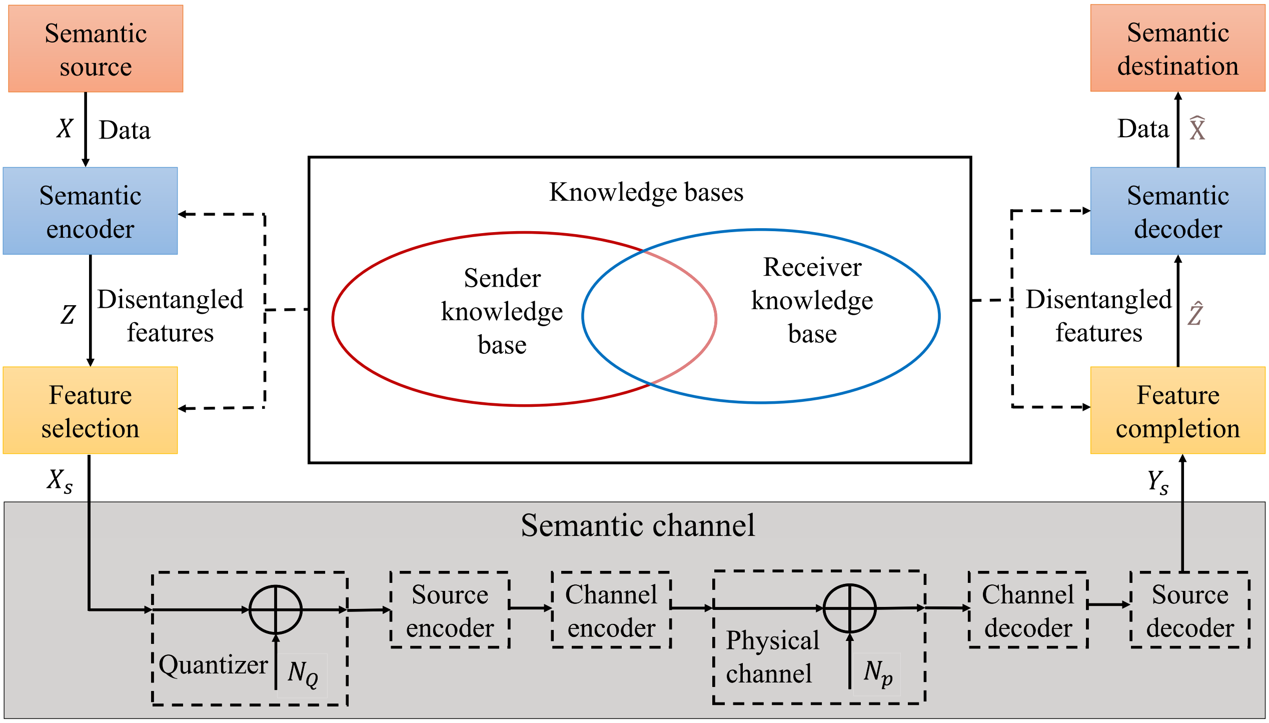


FIGURE 13. Task-oriented explainable SemCom – redrawn from [174, Fig. 1].

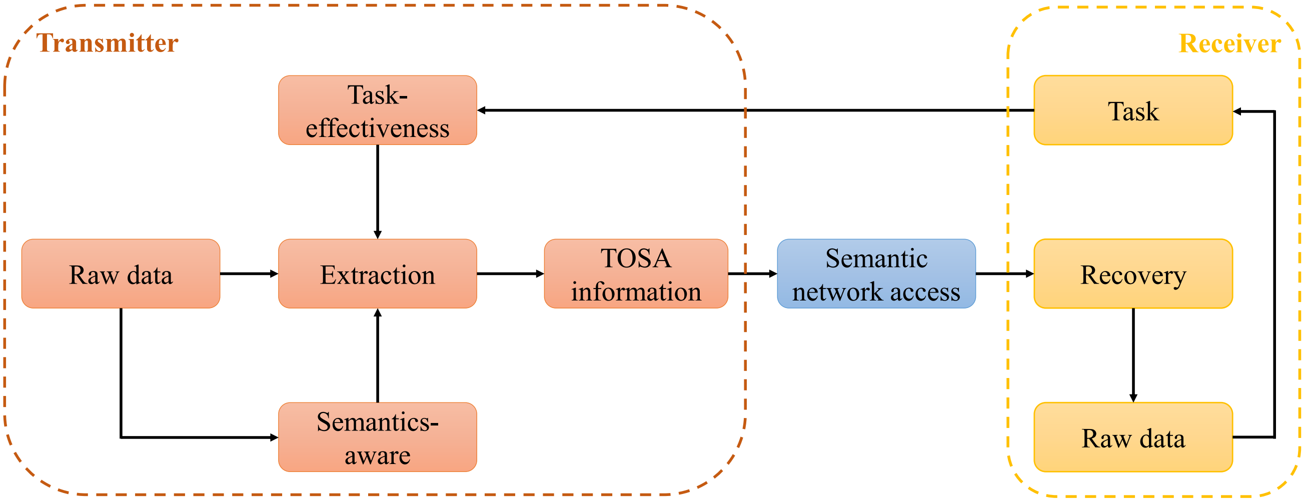


FIGURE 14. The TOSA framework proposed by the authors of [172] – redrawn from [172, Fig. 1].

with a TOSA recovery module. The recovery module produces the estimated raw data that will inform the execution of the task, whose effectiveness is fed to the task-effectiveness module at the transmitter. Per the task-effectiveness and semantics of the raw data, the TOSA information is extracted from the raw data, which can be text, speech, image, video, 360° video, haptic data, and sensor and control data.

For 1D data, such as speech and text, their semantic information can be extracted via *embedding* [172], which can be accomplished using state-of-the-art transformers such as BERT (bidirectional encoder representations from Transformers) [232]. For 2D data, such as an image, its edges, corners, blobs, and ridges can be extracted – using CNNs – as its semantic information [172]. For 3D data, such as video, its temporal correlation between the adjacent frames can be extracted as semantic information, also using CNNs [172]. For emerging XR applications, the 360° video is a new data type, with its semantic information being the *human field-of-view (FoV)* that occupies its almost 120° video. It has the highest resolution requirement at the center [172].

To extract FoV, biological information compression methods have been employed [172]. Moving to haptic data, they consist of two submodalities – *tactile information and kinesthetic* [172]. While the former is characterized by friction, hardness perception, warmth conductivity, macroscopic roughness, and microscopic roughness, the latter relates to the position/orientation of human body parts and external forces/torques applied to them [172]. To filter such haptic data that cannot be perceived by humans, *just noticeable difference (JND)* identifies the respective semantic information, which can be extracted by using *Weber's law* [172]. For sensor and control data, at last, their freshness is the corresponding semantic information that can be captured using *AoI (age of information)* [172].

The authors of [233] employed the TOSA framework for wireless UAV control and command downlink transmission. We now proceed with TOCD [88].

8) TOCD

To enhance the task effectiveness of CPS, the authors of [88] proposed a TOCD framework, as depicted in Fig. 15, which comprises the environment module, the multi-agent module, and the communication module. The communication module commands the communication capabilities and constraints (e.g., rate, power, energy, codeword length), which determines the nature of the inter-agent information exchange [88]. Incorporating actuator, sensor, integrated, and processor agents, the multi-agent module interacts with the environment module w.r.t. observations and probing actions [88]. The environment is defined by a set of parameters determining its state, constrained by the actuator and the actions of integrated agents. These actions, in turn, are translated into levels of task effectiveness [88]. Employing the task effectiveness values, the capabilities of the multi-agent module, and the constraints of the communication module, the proposed TOCD aims to optimize the multi-agent module by jointly designing its communication strategies and action policies [88].

To formally state the TOCD problem, let us briefly define the parameters of Fig. 15, considering a multi-agent system with K agents and the set $\mathcal{K} := \{1, 2, \ldots, K\}$. An agent k at the t-th time slot can observe the CPS state via the local observation signal $\boldsymbol{o}_k(t) = \boldsymbol{s}_k \in \mathcal{S}_k$, where $\mathcal{S}_k$ stands for the set of local system states [88]. Per such observations, the global system state at the t-th time is denoted as $\boldsymbol{s} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K] \in \mathcal{S}$, where $\mathcal{S} := \bigcup_{k \in \mathcal{K}} \mathcal{S}_k$ [88]. At any time slot t, the k-th agent can execute an action $\boldsymbol{p}_k(t)$, which can impact the overall state of the system, whose collective actions are represented by $\boldsymbol{p} := [\boldsymbol{p}_1(t), \boldsymbol{p}_2(t), \ldots, \boldsymbol{p}_K(t)] \in \mathcal{P}$ [88]. In connection with $\boldsymbol{s}$ and $\boldsymbol{p}$, let $\boldsymbol{s}(\cdot)$ and $\boldsymbol{p}(\cdot)$ be the sequence of system state and actions over time, respectively, that are the arguments of a *task effectiveness function* $T(\boldsymbol{s}(\cdot), \boldsymbol{p}(\cdot)) \in [0, 1]$ [88]. To make a decision on $\boldsymbol{p}$, one has to extract $\boldsymbol{c} := [\boldsymbol{c}_1(t), \boldsymbol{c}_2(t), \ldots, \boldsymbol{c}_K(t)] \in \mathcal{C}$ – contained in $\boldsymbol{s}$ – that is useful for the task at hand, where $\boldsymbol{c}_k(t)$ stands for the communication message sent by the k-th agent at the t-th time slot [88]. The communication network between agents is characterized through the mapping $h : \mathcal{C} \to \tilde{\mathcal{C}} : \boldsymbol{c} \mapsto \tilde{\boldsymbol{c}}$, where $\tilde{\boldsymbol{c}} := [\tilde{\boldsymbol{c}}_1(t), \tilde{\boldsymbol{c}}_2(t), \ldots, \tilde{\boldsymbol{c}}_K(t)]$ – given $\tilde{\boldsymbol{c}}_k(t)$ is the message received by the k-th agent at the t-th time slot. According to these definitions and parameters, the TOCD problem is stated below.

Definition 1 ([88, Definition 1]): To maximize the task effectiveness $T(\boldsymbol{s}(\cdot), \boldsymbol{p}(\cdot))$, TOCD aims to maximize the **communication policy** and **action policy**, as defined below.

- **Communication policy** $\pi^{(C)} : \mathcal{S} \to \mathcal{C} : \boldsymbol{s} \mapsto \boldsymbol{c}$. Because the extracted information from $\boldsymbol{s} \in \mathcal{S}$ needs to be transmitted to the decision maker(s) through the communication channel(s) $h : \mathcal{C} \to \tilde{\mathcal{C}}$, $\pi^{(C)}$ may include information distillation and source/channel coding policies [88].
- **Action policy** $\pi^{(P)} : \mathcal{S} \times \tilde{\mathcal{C}} \to \mathcal{P} : \boldsymbol{s} \times \tilde{\boldsymbol{c}} \mapsto \boldsymbol{p}$. Per the observed system states $\boldsymbol{s}$ and the communicated information $\tilde{\boldsymbol{c}}$, $\pi^{(P)}$ decides the action $\boldsymbol{p}$, which can be either a global action in the coordinator or the distributed actions in the local agents [88].

Compared with the conventional optimization problem that would search for an optimal policy by directly communicating states $\pi^* : \mathcal{S} \times \mathcal{S} \to \mathcal{P}$, TOCD has a two-fold advantage: *i*) reducing the cost of communication; *ii*) minimizing the complexity of optimizing the decision process [88]. Accordingly, TOCD has many applications in autonomous vehicles, federated learning, over-the-air computation for smart manufacturing plants, tactile Internet, 5G self-organized networks, industrial IoT, and UAV networks [88].

We now proceed with task-oriented ISCC in edge AI inference.

9) TASK-ORIENTED ISCC IN EDGE AI INFERENCE

At the confluence of edge computing and AI, *edge AI* [5], [51], [52], [62], [179], [184], [234], [235] has emerged as a promising 6G technology enabler, which employs well-trained ML models – at the edge – whose inference is used to make decisions. Such edge AI model inference or *edge inference* enables many 6G use cases, such as Metaverse and autonomous driving, inspiring the development of *on-device inference*, *on-server inference*, and *split inference* algorithms [173]. While on-device inference implements the inference task on a resource-constrained edge device, requiring heavy storage and computational cost, on-server inference implements the inference task at an edge server that receives the input data from edge devices, leading to data privacy leakage. To tackle these challenges, researchers propose split inference, in which the AI model is split into two sub-models that are deployed at the edge devices and edge server for feature extraction and inference, respectively [173].

In view of split inference, the ISCC framework in edge AI inference is schematized in Fig. 16, where the accuracy of split inference depends on data acquisition (sensing), feature extraction and quantization (computation), and feature transmission to an edge server (communication)

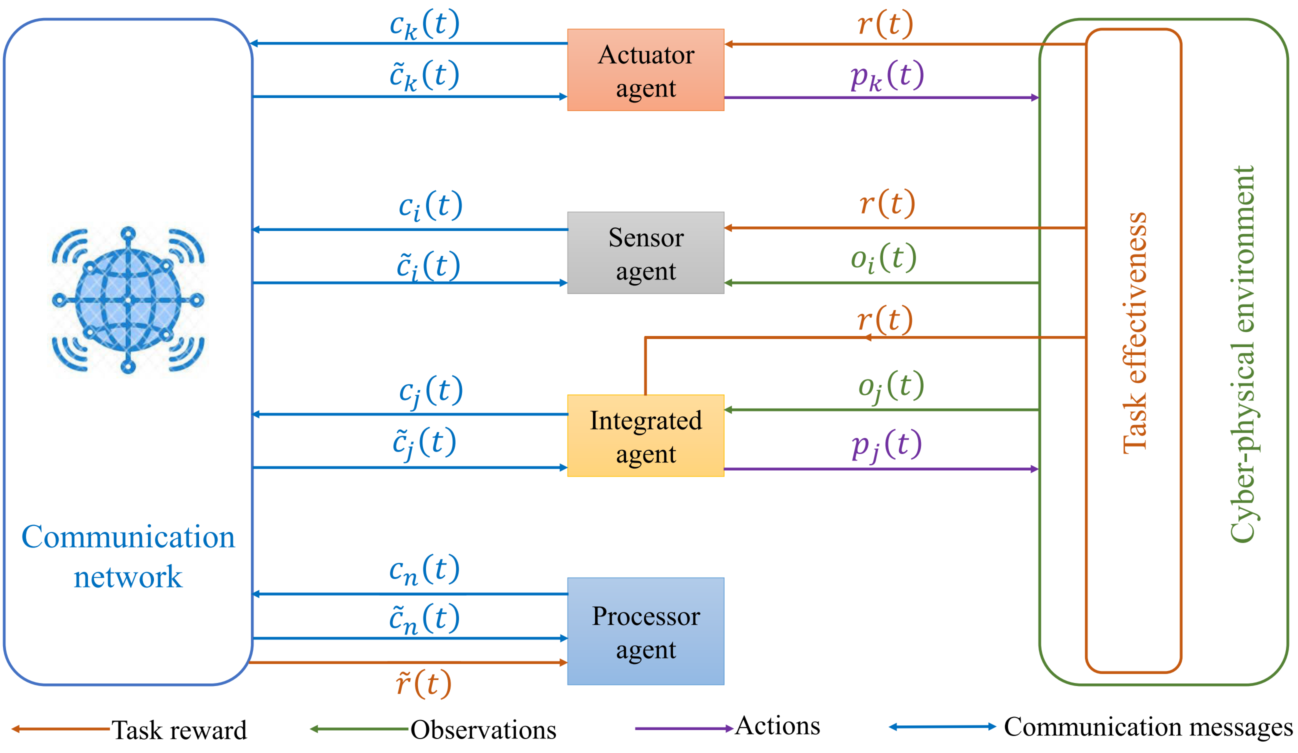


FIGURE 15. The TOCD framework proposed by the authors of [88] for cyber-physical systems – redrawn from [88, Fig. 3].

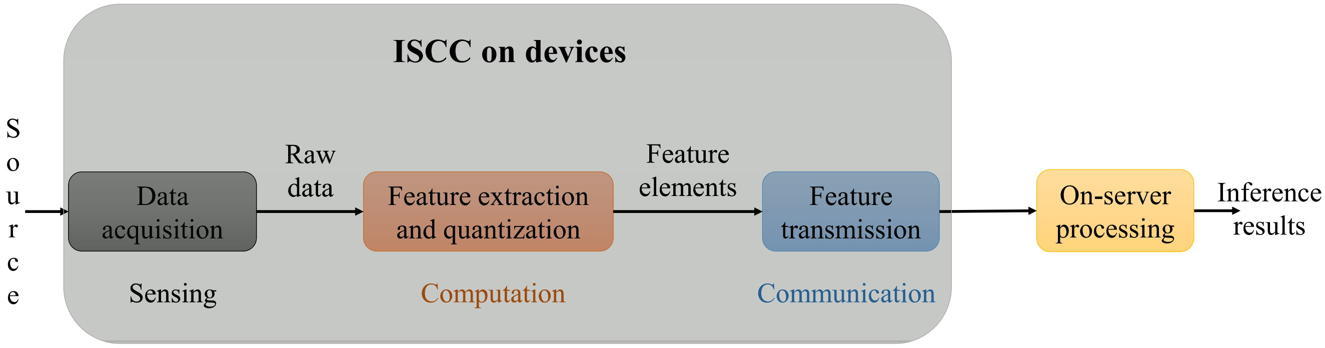


FIGURE 16. The ISCC framework in edge AI inference proposed by the authors of [173] – redrawn from [173, Fig. 1].

[173]. These three processes need to be integrated and jointly designed because of the following intertwined reasons: *i*) Sensing and communication compete for radio resources; *ii*) the permitted communication resource determines the required level of quantization; and *iii*) the quantized features must be reliably transmitted – under a delay constraint – to an edge server [173]. In addition, ISCC should be designed under a task-oriented principle that steers the successful completion of an inference task [173]. To tackle the challenges of task-oriented ISCC for multi-device edge AI, the authors of [173] formulated a non-convex inference accuracy maximization problem, under constraints of limited on-device resources and low-latency requirement, and proposed a *sum-of-ratios-based optimal ISCC scheme*. Evaluated and corroborated by simulations, this optimal scheme jointly determines the transmit power and time allocation at multiple sensing and communication devices, as well as their allocation of quantization bits for computation distortion control [173].

We now move on to major use cases of goal-oriented SemCom.

B. MAJOR USE CASES OF GOAL-ORIENTED SEMCOM

Similar to H2H SemCom, H2M SemCom, and M2M SemCom [76], the major use cases of goal-oriented SemCom can be classified as *H2H goal-oriented SemCom*, *H2M goal-oriented SemCom*, and *M2M goal-oriented SemCom*.[7] Major use cases of M2M goal-oriented SemCom include *autonomous transportation and autonomous vehicles*, *UAV communication networks*, *goal-oriented IoT*, *smart*

[7] Although the state-of-the-art goal-oriented SemCom literature comprises little to none information on the use cases of *H2H goal-oriented SemCom* and *H2M goal-oriented SemCom*, they are applicable for designing 6G H2H and H2M communication systems, respectively.

factories and IIoT, *NCSs*, *video conferencing*, and *Metaverse*, which are discussed below. We begin our discussion with autonomous transportation and autonomous vehicles.

1) AUTONOMOUS TRANSPORTATION AND AUTONOMOUS VEHICLES

Supporting the scalability of future massive networked intelligent systems, semantic-empowered communication will enhance network resource usage, energy consumption, and computational efficiency significantly, and thus pave the way for the design of next-generation real-time data networking [83]. This type of semantic networking will make it possible to transmit only informative data samples and convey only the information that is relevant, useful, and valuable for achieving its defined goals [83]. Accordingly, goal-oriented SemCom is a prime enabler for emerging 6G use cases, such as autonomous transportation and autonomous vehicles (AVs) [83].[8]

For a vehicular ad-hoc network of connected AVs, the authors of [236] proposed a goal-oriented SemCom framework, wherein a deep autoencoder (AE) is deployed to capture the semantic information from traffic signs. Specifically, the deep AE's encoder extracts – from a traffic sign – the respective semantic information, which is transmitted to the connected AVs [236]. The connected AVs then receive the transmitted information, which is then fed to their decoder (of the deep AE) in order to reconstruct the traffic sign [236]. Based on the reconstructed traffic sign, the connected AVs then deploy a DQN to take the proper action [236]. For this goal-oriented SemCom technique, the experimental results – reported by the authors of [236] – corroborate that the proposed scheme can drastically minimize the communication cost.

We now continue with UAV communication networks, as another major goal-oriented SemCom use case.

2) UAV COMMUNICATION NETWORKS

Continuing to play a pivotal role in 5G and beyond systems due to their flexible and low-cost deployments, UAVs often serve as a complementary application of the existing infrastructure to stand-alone service in emergency scenarios or remote areas [88]. A network of UAVs forms a multi-UAV system, where the main challenge – in contrast to a single-UAV system – lies in how to efficiently coordinate the UAVs' operations under limited communication resources [88]. In attempting to address this challenge, the state-of-the-art works on multi-UAV focus on the UAVs' action policy according to bit-rate maximization or packet-error rate minimization design objective [88]. Nevertheless, this communication system design is not optimal for a joint task at hand [88].

The TOCD framework (see Fig. 15) proposed by the authors of [88] – and discussed in Sec. IV-A8 – can be applied for designing K-UAV systems ($K \geq 2$), where each UAV observes the environment via its location (local state) $\boldsymbol{x}_k \in \mathcal{X}$ and receives message $\tilde{\boldsymbol{c}}_k \in \tilde{\mathcal{C}}$ from the centralized controller (or its neighbors) [88, Fig. 6]. Per [88, Fig. 6], specifically, each UAV can jointly optimize its action policy and communication message to be exchanged with the centralized controller (or neighboring UAVs), whose output is employed by each UAV to jointly determine its communication message $\boldsymbol{c}_k \in \mathcal{C}$ and next movement (action) $\boldsymbol{a}_k \in \mathcal{A}$ [88]. Representing all actions (movements) by $\boldsymbol{a} := [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$ and all locations of UAVs by $\boldsymbol{x} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K]$, the performance of the collaborative task in K-UAV systems can be resolved through a general utility function $\Phi(\boldsymbol{x}, \boldsymbol{a})$ [88]. Accordingly, each k-th UAV optimizes a joint communication and action policy $\pi_k^{(c,a)} : \mathcal{X} \times \tilde{\mathcal{C}} \rightarrow \mathcal{C}_k \times \mathcal{A}$ that maps the k-th local observation $\boldsymbol{x}_k$ and the k-th received message $\tilde{\boldsymbol{c}}_k$ to a tuple of the k-th local encoded message $\boldsymbol{c}_k$ and the k-th local movement $\boldsymbol{a}_k$ [88].[9]

We now continue with goal-oriented IoT.

3) GOAL-ORIENTED IOT

Enabled by a myriad of heterogeneous devices with sensing, actuation, processing, and wireless communication capabilities, IoT applications are fueling AI engines, which are in turn the cornerstone of IoT systems that lead to seamless integration of the interconnected physical world – of sensors, devices, and actions – and its programmable digital representation, forming a cyber-physical continuum with advanced intelligence and ubiquitous connectivity [169]. In view of this vision, while considering energy efficiency, communication overhead, computation capabilities, and so forth, it is essential to provide *distributed AI services* – with not only high reliability (e.g., high dependability and trustworthiness), but also ultra-low latency – for time-critical IoT applications, such as autonomous driving, industrial automation and control, and smart surveillance [169]. Driven by emerging 6G use cases such as virtual reality and autonomous driving, the current trend in IoT is marked by a growing demand for a wider bandwidth to accommodate the insatiable need for higher data rates, and elevated energy consumption [169]. Due to limited spectrum, energy availability, and computing power, pursuing these requirements will surely lead to a *performance bottleneck*, calling for the design and development of a new communication paradigm that can enable tomorrow's IoT applications [169].

Therefore, inspired by the emerging goal-oriented SemCom design paradigms, the authors of [169] propose a goal-oriented system design for the IoT, embracing the key aspects of architecture design and goal-oriented adaptive resource optimization. Specifically, they formally define a *communication goal* first as the fulfillment of a task with a target effectiveness level. They then pursue a joint

[8] For the same reason, goal-oriented SemCom can be the major enabler for consumer robotics, environmental monitoring, and telehealth [83]

[9] In view of [88, Fig. 6], the TOCD framework can also be applied to autonomous driving systems, while considering their larger dimensions of action and observation spaces, and more stringent requirements of task effectiveness [88].

system optimization encompassing sensing, communication, computation, learning, and control aspects, with the aim of attaining effective goal-oriented communications with a minimum cost, while striking the desired tradeoff between goal effectiveness and cost [169]. Leveraging the interplay between model-based optimization and purely data-driven approaches, the authors of [169] apply this goal-oriented system optimization to the following three use cases:

- *Goal-oriented compression for edge inference*: To perform real-time classification under latency and accuracy constraints, many IoT devices transmit compressed features [169].
- *Cooperative effective sensing*: Under mean-square error (MSE) constraints, the goal of the sensor network is to reconstruct a signal field by sending compressed and noisy data to an FC [169].
- *Goal-oriented federated learning*: By usually exchanging compressed models computed from locally collected data, a set of edge devices aims to train a common DL model – with a mediation of an edge server – under end-to-end delay and accuracy constraints [169].

We now proceed to our brief discussion of smart factories and IIoT.

4) SMART FACTORIES AND IIOT

In smart factories of the future, it will be crucial to limit the operation of machines to performing specific actions [73]. In this vein, goal-oriented SemCom can be designed and employed to convey only the semantic information of the control signals [73], so smart factories can reduce their communication cost and improve their operational efficiency [73]. In view of smart manufacturing, IIoT is the generic framework that leverages the availability of abundant data produced by sensors and various devices in order to enhance the accuracy, efficiency, and reliability of an industrial manufacturing process [88]. Through the availability of data generated by various devices and sensors, IIoT enables each manufacturing process to access a much more comprehensive view of the current state of the system [88]. However, due to the limitations in the communications rate in rural areas and/or the processing power of actuators and controllers, extracting useful information can be a challenge [88]. What can also be a challenge for the manufacturing value chain (i.e., inbound logistics, operations, and outbound logistics) is the many thousands of tiny elements generating several megabytes of data per second, making communication and processing power the bottlenecks of the IIoT system [88].

To mitigate the possible bottlenecks of the IIoT system, task-oriented communications facilitate the extraction of useful data generated by any controller/actuator of the system [88]. To this end, the TOCD framework schematized in Fig. 15 is also applicable for the IIoT systems, as schematized in [88, Fig. 5(a) and (b)] – i.e., [88, Fig. 5(a)] illustrates the IIoT problem for collocated plants and sensors; [88, Fig. 5(b)] for collocated controller and sensors.

We now continue with our brief discussion of NCSs.

5) NCSS

Emerging and futuristic NCSs are major use cases of goal-oriented SemCom, which require the joint optimization of the communication and control objectives [195]. State-of-the-art communication technologies, on the other hand, are agnostic to control objectives and pursue communication network and control system optimization separately, which is likely to yield suboptimal solutions by narrowing the solution space of the problem in both areas [195]. Accordingly, massive-scale NCSs can be enabled by unifying control and communication techniques under the umbrella of semantics of information. Therefore, fundamentally re-designing the techniques for information generation, transmission, transport, and reconstruction to optimize the performance of NCSs applications that utilize this information is of paramount importance [195].

Toward a goal-oriented fundamental redesign of NCSs, the first step would be distinguishing the tradeoff regions – between control and information costs – that shall be used as guidelines to design policies for scheduling data packets, while obeying processing and communication delays in observation/command channels [195]. It is also crucial to address the joint design problem in *hierarchical NCSs* under various information structures, and establish optimal decision-making points in the respective hierarchy [195]. In this vein, the high-level decision makers have more information – in contrast to low-level ones – about the whole system, though communicating with them incurs larger delays and is costlier [195].

We now move on to discuss video conferencing, as an emerging goal-oriented SemCom use case.

6) VIDEO CONFERENCING

During and after the COVID-19 pandemic, a popular mode of communication has been video conferencing, whose state-of-the-art technology relies on video compression that causes reduced resolution under a limited bandwidth [237]. To overcome this challenge, *semantic video conferencing (SVC)* maintains a high resolution by transmitting only some keypoints that capture the speakers' motions, since the background is often static and the speakers do hardly change, though the impact of transmission errors on keypoints has hardly been investigated [237]. Consequently, the authors of [237] put forward a *keypoint transmission-based SVC network* [237, Fig. 1], which can substantially minimize transmission resources, while only missing detailed expressions. In the proposed SVC network, transmission errors only lead to a changed expression, unlike in the conventional methods that directly destroy pixels, as described in the SVC framework below [237].

As depicted in [237, Fig. 1], the SVC framework comprises three levels: effectiveness, semantic, and technical levels, which all have a function in the SVC's goal of minimizing the difference between the transmitted and recovered frames [237].

- *The effectiveness level*: It delivers the motion and expression of the speaker [237].

- *The semantic level*: The speaker's photo (with a distinct face) is shared in advance, given the speaker has no considerable change during the video conversation, and the first video frame is shared – for convenience – with the receiver [237].
- *The technical level*: At the transmitter, the keypoint detector extracts the facial movement in the current frame, whose corresponding keypoints are transmitted at the technical level [237].

According to the received keypoints and the shared photo, the semantic module of the receiver reconstructs – employing the outputs of the technical level networks that are trained to cope with distortion and interference from wireless physical channels – the video frame [237]. Summing up, SVC has three subnets: A keypoint detector and a generator at the semantic level, and a dense layer-based encoder and decoder at the technical level [237]. Meanwhile, considering the effect of feedback in SVC, the authors of [237] devise another goal-oriented SemCom scheme dubbed an *SVC hybrid automatic repeat request (SVC-HARQ)*, with acknowledgment feedback for time-varying channels, while incorporating a semantic error detector as in [237, Fig. 2]. Moreover, considering CSI feedback, the authors of [237] enhance SVC by proposing SVC-CSI, which learns to allocate more information at subchannels with higher channel gains [237].

Goal-oriented SemCom also enables haptic communication [238], [239], [240], [241]; tactile Internet [242], [243], [244]; and digital twins [162], [163], [164], [165], [245], [246] that comprise the *Metaverse engine*, which enables the Metaverse, as discussed below.

7) METAVERSE

Let us begin by providing some representative definitions of the Metaverse.

Definition 2: "The metaverse represents the top-level hierarchy of persistent virtual spaces that may also interpolate in real life, so that social, commercial, and personal experiences emerge through Web 3.0 technologies" [247].

Definition 3: "The Metaverse is an embodied version of the Internet that comprises a seamless integration of interoperable, immersive, and shared virtual ecosystems navigable by user-controlled avatars" [32].

Definition 4: "The metaverse is a computer-generated world with a consistent value system and an independent economic system linked to the physical world" [170].

Definition 5: "As a new Internet application, Metaverse integrates a variety of new technologies and has the characteristics of multi-technology; as a new social form, Metaverse has the characteristics of sociality; as a parallel and closely related to the real world in the virtual world, Metaverse has the characteristics of hyper spatiotemporality" [248].

Definition 6: "Metaverse is a compound word of transcendence meta and universe and refers to a three-dimensional virtual world where avatars engage in political, economic, social, and cultural activities" [249].

Definition 7: "In this paper, we consider the metaverse as a virtual environment blending physical and digital, facilitated by the convergence between the Internet and Web technologies, and extended reality (XR). To achieve such duality, the development of metaverse has to go through three sequential stages, namely (I) *digital twins*, (II) *digital natives*, and eventually (III) *co-existence of physical-virtual reality* or namely the surreality" [250].

Definition 8: "At present, the metaverse is defined as a shared virtual 3D world or even multiple cross-platform worlds that can provide users a comprehensively immersive experience with interactive and collaborative activities. Besides virtual places and constructions fixed in the virtual world, many other entities, such as objects, user identities, and digital goods, can be exchanged between different virtual worlds and even reflected into the reality world" [251].

Definition 9: "As a new type of Internet application, it aims to build a virtual world parallel to the real world with a stable society and economic system, allowing each user to produce content and edit the world. Some call it the new form of the Internet, some call it the digital layer of everyday life, and still others call it a fusion of virtual and physical reality, a persistent virtual space, or a digital twin of the real world. The Metaverse is still an evolving concept, with different players enriching its meaning in their own way" [252].

The Metaverse integrates a suite of seven underlying technologies, namely *interactivity*, *digital twin*, *tactile internet*, *blockchain*, *AI*, *ubiquitous computing*, and *5G and beyond networking* [32], [170], [171]:

- *Interactivity*: It enables immersive experience via virtual reality (VR), augmented reality (AR), mixed reality (MR), XR, wearable sensors, and BCI [170].
- *Digital twin*: It produces high-fidelity simulated spaces through *3D simulation* and *3D modeling* [170].
- *Tactile internet*: It enables H2H and H2M communication.
- *Blockchain*: It allows the Metaverse economic systems and identification through *smart contracts*, *non-fungible tokens (NFTs)*, and *decentralized finance (De-Fi)* [170].
- *AI*: It facilitates big data inference and smart decision through computer vision, natural language processing, ML, DL, and RL [170].
- *Ubiquitous computing*: It enables data processing and storage via cloud-edge-end computing [170].
- *5G and beyond networking*: It enables ubiquitous sensing and transmission through IoT, software-defined networking (SDN), and beyond 5G/6G [170].[10]

As schematized in Fig. 17, the Metaverse relies on an infrastructure for sensing & actuation, communication & networking, computation, and storage, which feed big data to *the Metaverse engine*. Enabled by core technologies for interactivity, tactile internet, digital twin, AI, and the

[10] The Metaverse requires a multi-sensory multimedia network, which can be designed using one or more of the following mathematical frameworks: *Optimization theory*, *auction theory*, *contract theory*, *machine learning*, *stochastic geometry*, and *game theory* [253].

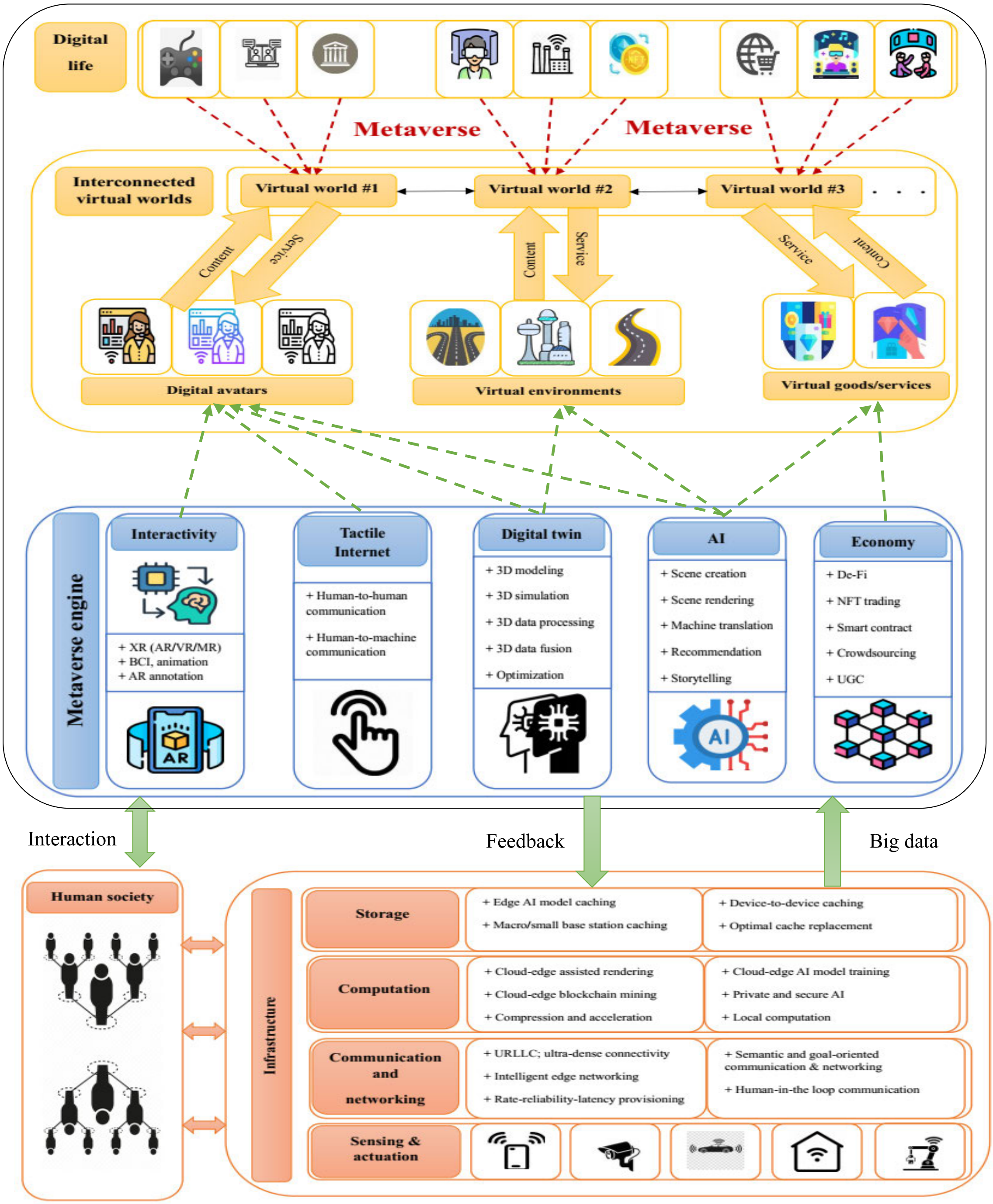


FIGURE 17. Metaverse architecture in view of integrating the human, the physical, and the digital worlds – drawn by unifying [170, Fig. 5], [32, Fig. 1], and [171, Fig. 1]. XR, AR, VR, MR, De-Fi, NFT, and UGC abbreviate extended reality, augmented reality, virtual reality, mixed reality, decentralized finance, non-fungible token, and user-generated content, respectively.

Metaverse economy, the Metaverse engine pushes multisensory multimedia content to the Metaverse (interconnected virtual worlds), while also bringing feedback to the human users through the Metaverse infrastructure. Through their digital avatars, which are the digital representation of human users in the Metaverse [170], users can interact with the Metaverse by requesting services from one or more virtual worlds. Specifically, users in the Metaverse can access virtual environments that are the simulated real or imaginary environments (consisting of 3D digital things and their attributes) in the Metaverse; sell/buy virtual goods that are tradeable commodities (e.g., digital arts and land parcels) produced by virtual service providers or the Metaverse users; and sell/buy virtual services that broadly include digital market, digital currency, digital regulation, social service, etc. [170].

In view of Fig. 17, the following are the six pillars of the Metaverse ecosystem: *Avatar*, *content creation*, *virtual economy*, *social acceptability*, *security & privacy*, and *trust & accountability* [250]. In view of these pillars, here are the eight pillars of Metaverse technology enablers: *network*, *edge/cloud*, *AI*, *computer vision*, *blockchain*, *robotics/IoT*, *user interactivity*, and *XR* [250]. In line with Fig. 17, meanwhile, the seven inside-to-outside layers of the Metaverse are [251, Fig. 3]: *infrastructure*, *human interface*, *decentralization*, *spatial computing*, *creator economy*, *discovery*, and *experience* [251]. Accordingly, the *key requirements* of the Metaverse are outlined below.

- The Metaverse must have *persistence*, *immersiveness and realism*, *multimodal interaction*, *security*, and *social interaction* [254].
- The Metaverse should have *connectivity*, *decentralization*, *AI*, *interoperability*, and *openness* [254].
- The Metaverse would be good to have *scalability*, *configurability*, *privacy and ethics*, *accessibility*, and *market access* [254].

Generalizing the Metaverse vision per Fig. 17, the authors of [255] envision the Metaverse as an intersection of seven worlds and experiences, namely, *physical, digital, and virtual worlds – with cyber, extended, live, and parallel experiences* [255, Fig. 2], which have unique challenges and applications. The Metaverse has many broad applications in *education*; *smart city and smart home* (e.g., city planning, designing and planning, intelligent building, smart transportation, IoT maintenance, and environmental protection); *entertainment* (massively multiplayer online game); *modeling, monitoring, and warning*; *autonomous driving*; *culture* (tourism, museum, and art exhibition); *medicine* (telemedicine, medical education, and healthcare); *business* (retail, marketing, virtual assistant, recommendation system, and immersive business); *real estate*; *socialization*; and *manufacturing* (product design collaboration, production processes' optimization, and increased transparency for customers) [252], [256], [257].

Several types of Metaverse have been envisioned to date: *Edge-enabled Metaverse* [32], [171]; *semantic Multiverse* [258]; *ubiquitous semantic Metaverse* [259]; *SemCom-assisted Metaverse* [260]; *ubiquitous semantic Metaverse* [257]; *decentralized Metaverse* [245]; and *quantum-enabled wireless Metaverse* [261]. Meanwhile, enabling the wireless Metaverse comes down to enabling realistic multi-modal interactions among a set of human and machine agents in a vast range of mobile scenarios [258]. Because both SemCom and goal-oriented SemCom employs only semantically-relevant information, while minimizing bandwidth consumption, power usage, and transmission delay, they are highly useful for realizing the wireless Metaverse. In this vein, to unify SemCom and *AI-generated content (AIGC)* [262] in the Metaverse, the authors of [263] put forth a new framework termed integrated SemCom and AIGC (ISGC). ISGC can efficiently extract semantic information, produce high-quality content with AI, and seamlessly integrate the AIGC into the Metaverse ecosystem, while unifying SemCom, AIGC, and the Metaverse [263].

In regards to the sensing and actuation infrastructure layer of the Metaverse, IoT and sensor networks installed in the physical world collect a vast amount of data from the environment [171]. To manage such big data, both SemCom and goal-oriented SemCom are of paramount importance by transmitting only semantically-relevant information, while minimizing bandwidth consumption, power usage, and transmission delay. Advocating this paradigm shift for the Metaverse, akin to several the state-of-the-art Metaverse works, the authors of [256] put forward a *task-oriented Metaverse design*, by also introducing the Metaverse's three infrastructure pillars (i.e., human-computer interface, sensing and communications, and network architecture), and envisioning the roadmap toward the full vision of the Metaverse. The authors of [264] devise a task-oriented and semantics-aware communication framework designed to improve the effectiveness and efficiency of avatar-based communication in wireless AR applications. By introducing new semantic information whose relationships are represented by a graph, this framework extracts and transmits only essential semantic information in wireless AR communication, considerably reducing bandwidth consumption [264].

As one of the Metaverse technology enablers, the conventional immersive XR environment creation encompasses the following five steps (with their respective outputs): *calibration (single-frame point cloud)*, *registration (registered point cloud)*, *volume reconstruction (voxel frame)*, *marching cubes (meshes)*, and *rendering (3D environment)* [265, Fig. 1]. These steps are traditionally executed at a given location before transmitting the generated 3D environment to a remote user [265], rendering this technology *ultra-high-data-rate hungry*. To address this challenge and reduce end-to-end latency by effectively compressing the extensive XR data volume, the authors of [265] devise a semantic compression technique [265, Fig. 3] that splits the traditional volume reconstruction of [265, Fig. 1] into a *client-side virtual network function (VNF)* and a *server-side VNF* that semantically compress the registered point cloud into *color codes (CCs)* for transmission over the network and decode the received CCs to obtain the voxel frames, respectively. For this proposed scheme, the authors of [265] demonstrate – interestingly – that its aggregate latency is approximately one fifth of the transmission latency of the raw point cloud data, one third of the aggregate latency of the JPG/PNG compressed point cloud data, and one tenth of the aggregate latency of a networked XR that semantically compresses – directly – the camera-captured color and depth data (the benchmark). In the spirit of the work in [265] and for point cloud video data (a volumetric data format), the authors of [266] propose an AI-powered and semantic-aware transmission scheme – which considerably minimizes the transmitted data volume – that extracts semantic features from raw point cloud for efficient transmission, and then executes point cloud reconstruction. This end-to-end design eliminates the need for various conventional video transmission schemes, such as compression and codec [266].

Despite its limitless possibilities as an emerging technology [247], the Metaverse faces multi-faceted security threats from the following seven dimensions: authentication & access control, data management, privacy-related, network-related, economy-related, physical/social effects, and governance-related [170], [253]. In mitigating some of these threats, SemCom and goal-oriented SemCom can help, since semantic decoding hinges on the availability of destination KB that is shared with the source KB in real time [267]. Summing up, goal-oriented SemCom also has many other applications and use cases, including fault detection [268].

We now move on to the major state-of-the-art mathematical frameworks of goal-oriented SemCom.

V. MAJOR STATE-OF-THE-ART MATHEMATICAL FRAMEWORKS OF GOAL-ORIENTED SEMCOM

State-of-the-art goal-oriented SemCom algorithms were developed using the tools of information theory [269], IB, and variants of IB such as *robust IB* [87] and *distributed IB* [210]. These state-of-the-art frameworks of goal-oriented SemCom are briefly discussed henceforward, beginning from the basics of information theory.

A. BASICS OF INFORMATION THEORY

To underscore the basics of information theory, we hereinafter provide a brief discussion on *entropy*, *conditional entropy*, *relative entropy and mutual information*, and *conditional mutual information*, beginning with entropy.

1) ON ENTROPY

We start by defining the entropy of a discrete RV.

Definition 10: For a discrete RV X, its entropy $H(X)$ is defined by [269, eq. (2.1)]

$$H(X) := -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x), \tag{1}$$

where $\mathcal{X}$ is the alphabet, $p(x) := p_X(x) = \mathbb{P}(\{X = x\})$ is the PMF of X, and $H(X)$ is quantified in bits [269].

The entropy defined in (1) is often referred to as *Shannon entropy*, and $H(X) \geq 0$ [269, Lemma 2.1.1]. Meanwhile, if $X \sim p(x)$, the expected value of an RV $g(X)$ is equated as [269, eq. (2.2)]

$$\mathbb{E}\{g(X)\} := \sum_{x\in\mathcal{X}} g(x)p(x). \tag{2}$$

Thus, it follows from (2) and (1) that

$$H(X) = -\mathbb{E}\{\log_2 p(X)\} = \mathbb{E}\{1/\log_2 p(X)\}. \tag{3}$$

As a generalization of the entropy definitions in (3) and (1), we provide below the definition of *joint entropy*.

Definition 11: For a pair of discrete RVs (X, Y) with a joint PMF $p(x, y)$, their joint entropy $H(X, Y)$ is defined as [269, eq. (2.8)]

$$H(X, Y) := -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x, y), \tag{4}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the alphabets of X and Y, respectively, and $p(x, y) := P_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$.
To express the right-hand side (RHS) of (4) using expectation, we provide the following definition of the expectation of a function of multi-variate RVs: if $X \sim p(x)$ and $Y \sim p(y)$, the expected value of an RV $g(X, Y)$ takes the form [270]

$$\mathbb{E}\{g(X, Y)\} := \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g(x, y)p(x, y). \tag{5}$$

Thus, using (5), the joint entropy – as it is defined in (4) – can also be expressed as [269, eq. (2.9)]

$$H(X, Y) = -\mathbb{E}\{\log_2 p(X, Y)\}. \tag{6}$$

We now move on to highlight the conditional entropy.

2) ON CONDITIONAL ENTROPY

Underneath, we define the *conditional entropy* of an RV given another RV.

Definition 12: For a pair of discrete RVs $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y|X)$ is defined as [269, eq. (2.10)]

$$H(Y|X) := \sum_{x\in\mathcal{X}} p(x)H(Y|X = x). \tag{7}$$

The RHS of (7) can then be simplified as

$$\begin{aligned}
H(Y|X) &\overset{(a)}{=} -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log_2 p(y|x) &\text{(8a)}\\
&\overset{(b)}{=} -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y|x)\log_2 p(y|x) &\text{(8b)}\\
&\overset{(c)}{=} -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(y|x) &\text{(8c)}\\
&\overset{(d)}{=} -\mathbb{E}\{\log_2 p(Y|X)\}, &\text{(8d)}
\end{aligned}$$

where (a) is due to the entropy definition in (1), (b) follows from rearranging the RHS of (8a), (c) is for the definition of the conditional PMF $p(y|x)$ with $p(y|x) := p_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x) = p(x, y)/p(x)$ [270], and (d) is because of the definition in (5). It is intuitive from (8d) that $H(Y|X) \neq H(X|Y)$ [269].

If we now simply add (8d) and (6), it follows that

$$\begin{aligned} &H(Y|X)+H(X,Y) \\ &\quad = -\big[\mathbb{E}\{\log_2 p(Y|X)\} + \mathbb{E}\{\log_2 p(X,Y)\}\big] \\ &\quad \overset{(a)}{=} -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)[\log_2 p(y|x) \\ &\qquad + \log_2 p(x,y)] \\ &\quad \overset{(b)}{=} -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)[\log_2 p(y|x)p(x,y)] \\ &\quad \overset{(c)}{=} -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)[\log_2 [p(x,y)]^2/p(x)], \end{aligned} \tag{9}$$

where (a) is due to (8c) and (4), (b) is due to the property of the logarithm, and (c) is for $p(y|x) := p(x,y)/p(x)$ [270]. Applying the properties of logarithm to the RHS of the equality (c) leads to

$$\begin{aligned} &H(Y|X)+H(X,Y) \\ &\quad = -2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y) \\ &\qquad + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x) \\ &\quad \overset{(e)}{=} -2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y) \\ &\qquad + \sum_{x\in\mathcal{X}} p(x)\log_2 p(x) \overset{(f)}{=} 2H(X,Y) - H(X), \end{aligned} \tag{10}$$

where (e) is due to the property of the joint PMF $p(x,y)$ with $p(x) = \sum_{y\in\mathcal{Y}} p(x,y)$ [270], and (f) follows from (4) and (1). Rearranging (10) gives the result

$$H(Y|X) + H(X) = H(X,Y). \tag{11}$$

This is an important result that is widely known as the *chain rule* [269] and formalized below.

Theorem 1 (Chain rule [269, Th. 2.2.1]): For a pair of discrete RVs $(X,Y) \sim p(x,y)$,

$$H(X,Y) = H(X) + H(Y|X). \tag{12}$$

From (12), the following corollary [269, eq. (2.21)] follows.

Corollary 1:

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z). \tag{13}$$

We now proceed to highlight relative entropy and mutual information.

3) ON RELATIVE ENTROPY AND MUTUAL INFORMATION

The relative entropy between two PMFs is defined below.

Definition 13: The relative entropy or Kullback-Leibler (KL) distance between two PMFs $p(x)$ and $q(x)$ is given[11] by [269, eqs. (2.26) and (2.27)]

$$D(p||q) := \sum_{x\in\mathcal{X}} p(x)\log_2\frac{p(x)}{q(x)} = \mathbb{E}_{p(x)}\Big\{\log_2\frac{p(X)}{q(X)}\Big\}, \tag{14}$$

where the conventions $0\log_2\frac{0}{0} = 0$, $0\log_2\frac{0}{q} = 0$, and $p\log_2\frac{p}{0} = \infty$ are used [269].

W.r.t. the relative entropy defined in Definition 13, the *mutual information* between two RVs is defined below.

Definition 14: For two discrete RVs X and Y with a joint PMF $p(x,y)$ and marginal PMFs $p(x)$ and $p(y)$, respectively, their mutual information $I(X;Y)$ is the relative entropy between $p(x,y)$ and the product distribution $p(x)p(y)$ [269, eqs. (2.28)–(2.30)]:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \tag{15a}$$

$$= D(p(x,y)||p(x)p(y)) \tag{15b}$$

$$= \mathbb{E}_{p(x,y)}\Big\{\log_2\frac{p(X,Y)}{p(X)p(Y)}\Big\}. \tag{15c}$$

From (15c), it directly follows that

$$I(Y;X) = \mathbb{E}_{p(y,x)}\Big\{\log_2\frac{p(Y,X)}{p(Y)p(X)}\Big\} = I(X;Y). \tag{16}$$

The equality in (16) states the *symmetrical* nature of mutual information: i.e., X says as much about Y as Y says about X [269]. Meanwhile, simplifying the RHS of (15a), the mutual information $I(X;Y)$ can also be expressed as [269, eq. (2.39)]

$$I(X;Y) = H(X) - H(X|Y). \tag{17}$$

Thus, it follows from (17) and (16) that

$$I(X;Y) = \overbrace{H(Y) - H(Y|X)}^{=I(Y;X)}. \tag{18}$$

From the chain rule as expressed in (12), $H(Y|X) = H(X,Y) - H(X)$. Substituting this inequality into the RHS of (18) leads to the relationship [269, eq. (2.41)]

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{19}$$

At last, we note that [269, eq. (2.42)]

$$I(X;X) = H(X) - H(X|X) \overset{(a)}{=} H(X) - 0 = H(X), \tag{20}$$

where (a) follows through Definition 12 and (8c) due to the probabilistic fact[12] that $p(x|x) = 1$ for all $x \in \mathcal{X}$. In summary, the information-theoretic results in (16)-(20) are formalized in the following theorem.

Theorem 2 (Mutual information and entropy [269, Theorem 2.4.1]): The underneath results [269, eqs. (2.43)–(2.47)] are valid concerning the relationship

[11] Throughout this paper, we follow definitions w.r.t. the logarithm to the base two. However, the logarithm to the base ten are generally used, as it is also the case with some of the literature [269].

[12] Intuitively, $H(X|X) = 0$ is the reflection of the fact that there is no any uncertainty about $x \in \mathcal{X}$ provided that x is already known/given.

between mutual information and entropy:

$$I(X;Y) = H(X) - H(X|Y) \tag{21a}$$
$$I(X;Y) = H(Y) - H(Y|X) \tag{21b}$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{21c}$$
$$I(X;Y) = I(Y;X) \tag{21d}$$
$$I(X;X) = H(X). \tag{21e}$$

We build on the chain rule as stated in (12) and continue with the chain rules for entropy and mutual information. Beginning with the former, we state the following chain rule for the entropy of a collection of RVs.

Theorem 3 (Chain rule for the entropy of a collection of RVs [269, Th. 2.5.1]): For discrete RVs $X_1, X_2, \ldots, X_n$ drawn according to $p(x_1, x_2, \ldots, x_n)$, their joint entropy $H(X_1, X_2, \ldots, X_n)$ can be expressed as [269, eq. (2.48)]

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \tag{22}$$

Proof. The proof is provided in [269, p. 22-23].

We now proceed with our brief discussion on conditional mutual information.

4) ON CONDITIONAL MUTUAL INFORMATION

We define *conditional mutual information* below [269].

Definition 15 For discrete RVs $X, Y, Z \sim p(x, y, z)$, the conditional mutual information of X and Y given Z is defined by [269, eqs. (2.60) and (2.61)]

$$I(X;Y|Z) := H(X|Z) - H(X|Y,Z) \tag{23a}$$
$$= \mathbb{E}_{p(x,y,z)}\left\{ \log_2 \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \right\}. \tag{23b}$$

Definition 15 and Theorem 3 then lead to the following theorem on the chain rule for mutual information.

Theorem 4 (Chain rule for mutual information [269, Th. 2.5.2]): The following result is valid for the mutual information of multiple RVs [269, eq. (2.62)]:

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1). \tag{24}$$

Proof. Theorem 4 follows from Theorem 2 and (21a) that

$$\begin{aligned} I(X_1, X_2, \ldots, X_n; Y) &= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n|Y) \\ &\overset{(a)}{=} \sum_{i=1}^{n} \big[H(X_i|X_{i-1}, \ldots, X_1) \\ &\quad - H(X_i|X_{i-1}, \ldots, X_1, Y) \big] \\ &\overset{(b)}{=} \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1), \end{aligned} \tag{25}$$

where (a) is due to (22) and (b) follows from (23a). The last equation on the RHS of (25) is the RHS of (24). This completes the proof of Theorem 4. ■

We now move on to the basics of IB.

B. BASICS OF INFORMATION BOTTLENECK (IB)

Assume a source encoding of an information source is denoted by an RV X and we wish to obtain its relevant quantization $\tilde{X}$ to compress X as much as possible. Assume also that a relevance RV denoted by Y (e.g., a classification label) that must not be independent from X [208]. Thus, X and Y, have a positive mutual information $I(X;Y)$, and we presume that we have access to the joint PDF $p(x, y)$ [208], [271]. Nonetheless, under these settings and contrary to the rate-distortion problem, we would like $\tilde{X}$ (the quantized information) to capture as much information about Y (the relevance RV) as possible [208]. The amount of information about Y that is in $\tilde{X}$ is given by $I(\tilde{X};Y)$ and defined as [208, eq. (14)]

$$I(\tilde{X};Y) = \sum_{y} \sum_{\tilde{x}} p(y, \tilde{x}) \log \frac{p(y, \tilde{x})}{p(y)p(\tilde{x})} \overset{(a)}{\leq} I(X;Y), \tag{26}$$

where (a) is because lossy compression cannot convey more information than the original signal, and hence, there is always a tradeoff between rate and distortion [208]. Similarly to rate and distortion, there is a natural tradeoff between preserving meaningful information and compressing the original signal [208]. Bearing in mind this tradeoff, the IB problem concerns maintaining a constant amount of meaningful information about the relevant signal Y whilst minimizing the number of bits from the original information source X (maximizing its compression) [208]. This is equivalent to maximizing the meaningful information for a fixed compression of the original information signal [208]. Accordingly this amounts to passing the information that X provides about Y via a "bottleneck" formed by the compact information content in $\tilde{X}$ [208].

On par with the aforementioned motivation, the IB problem boils down to solving the following optimization problem [208, eq. (15)], [76, eq. (2)]:

$$\min_{p(\tilde{x}|x)} I(\tilde{X};X) - \beta I(\tilde{X};Y), \tag{27}$$

where the conditional distribution $p(\tilde{x}|x)$ represents the considered source encoder and β denotes the Lagrange multiplier connected to the constrained meaningful information [76], [208]. Meanwhile, the optimal solution for (27) – i.e., the optimal source encoder – is task-dependent, and a generic algorithm computes the optimal solution by alternating iterations [76]. In every iteration, minimization is performed by converging alternating iterations w.r.t. the PDFs $p(\tilde{x}|x)$, $p(\tilde{x})$, and $p(y|\tilde{x})$ [271, Th. 5]. This IB approach provides a unified framework for various information processing problems, including prediction, filtering and learning [208]. Toward these ends, IB has many applications in DL [271], ML [272], SemCom [273], and goal-oriented SemCom [207].

We now proceed to the variants of IB.

C. VARIANTS OF IB

To inspire much more work on goal-oriented SemCom algorithms and theories, we highlight below the principles of graph IB (GIB) [274], robust IB (RIB), deterministic IB, and distributed IB (DIB), beginning with GIB.

1) GRAPH IB (GIB)

To formally define GIB, which is proposed by the authors of [274], let Y be the target, $\mathcal{D} := \{(\boldsymbol{A}, X)\}$ be the input data for $\boldsymbol{A}$ being the graph structure and X being the node features, and Z be the representation. Concerning Z being the representation, GIB is used to optimize Z to capture the minimal sufficient information in input data $\mathcal{D}$ to predict the target Y [274]. To this end, the GIB problem reduces to solving the following optimization problem [274]:

$$\min_{p(Z|\mathcal{D})\in\Omega} \mathrm{GIB}_{\beta}(\mathcal{D}, Y; Z) := [-I(Y;Z) + \beta I(\mathcal{D};Z)], \tag{28}$$

where Ω represents the search space of the optimal model $p(Z|\mathcal{D})$ [274].

We now continue with RIB.

2) ROBUST IB (RIB)

The authors of [87] propose to use a design criterion named RIB to design the goal-oriented SemCom system schematized in [87, Fig. 1]. To define RIB formally, let the RVs X, Y, Z, and $\hat{Z}$ be the input datum, the target (label), the output of an encoder modeled as $p_{\phi}(z|x)$, and the output of a demodulator, respectively. From the vantage point of data compression, the optimal Z can be approximated by optimizing the IB problem [87] such that $I(Y;\hat{Z})$ is maximized while being subjected to the constraint on the amount of preserved information $I(X;\hat{Z})$ [87, eq. (5)]:

$$\max_{p_{\phi}(z|x)} I(Y;\hat{Z}) - \beta I(X;\hat{Z}). \tag{29}$$

Apart from data compression, another crucial goal-oriented SemCom design criterion is the maximization of the transmission rate, and hence [87, eq. (6)]

$$\max_{p_{\phi}(z)} I(Z;\hat{Z}), \tag{30}$$

where $p_{\phi}(z)$ is the marginal distribution that depends on the parameters ϕ [87]. Meanwhile, combining (29) and (30) leads to the RIB design principle (or criterion) that is given by [87, eq. (7)]

$$\max_{p_{\phi}(z|x)} I(Y;\hat{Z}) + \beta[I(Z;\hat{Z}) - I(X;\hat{Z})], \tag{31}$$

where β is fixed and $\beta \geq 0$.

We now move on to highlight deterministic IB [212].

3) DETERMINISTIC IB

The authors of [212] introduce a modified IB criterion named deterministic IB, which they say better captures the essence of compression than an optimal tradeoff between discarding as many bits as possible and selectively keeping the ones that are most important [212]. Meanwhile, the deterministic IB problem boils down to solving the following optimization problem [212, eq. (8)]:

$$\min_{p(\tilde{x}|x)} H(\tilde{X}) - \beta I(\tilde{X};Y), \tag{32}$$

where the deterministic IB optimization of (32) is subjected to the Markov constraint $\tilde{X} \leftrightarrow X \leftrightarrow Y$ [212].

We now proceed to emphasize DIB [212].

4) DISTRIBUTED IB (DIB)

To state and discuss the DIB framework [210], we must first consider the distributed learning (e.g., multi-view learning) model depicted in [210, Fig. 1]. Per [210, Fig. 1], Y is the signal to be predicted and $(X_1, \ldots, X_K)$ are the relevant K views of Y that could each be useful to understand one or more aspects of it [210]. Accordingly, the relevant observations could be either distinct or redundant. This justifies the assumption $(X_1, \ldots, X_K)$ are independent given Y [210]. This distributed learning problem's problem formulation [210, Sec. 2] is highlighted below.

Let $K \in \mathbb{N}_{\geq 2}$ be given and $\mathcal{K} := [K]$. Let $(X_1, \ldots, X_K, Y)$ be a tuple of RVs that have a joint PMF $p_{X_{\mathcal{K}},Y}(x_{\mathcal{K}}, y) := p_{X_1,\ldots,X_K,Y}(x_1, \ldots, x_K, y)$ for $(x_1, \ldots, x_K) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_K$ and $y \in \mathcal{Y}$, given that $\mathcal{X}_k$ for all $k \in \mathcal{K}$ and $\mathcal{Y}$ represent the alphabet of X_k and Y, respectively. Meanwhile, the Markov chain below is assumed to hold for all $k \in \mathcal{K}$ [210, eq. (3)]:

$$X_k \leftrightarrow Y \leftrightarrow X_{\mathcal{K}/k}, \tag{33}$$

i.e., $p(x_{\mathcal{K}}, y) = p(y)\prod_{k=1}^{K} p(x_k|y)$ for $x_k \in \mathcal{X}_K$ and $y \in \mathcal{Y}$. The distributed learning problem aims to characterize how the goal variable Y can be accurately estimated from the observations $(X_1, \ldots, X_K)$ when they are processed individually in different encoders [210].

Moreover, let a training dataset $\{(X_{1,i}, \ldots, X_{K,i}, Y_i)\}_{i=1}^{n}$ comprise n independent and identically distributed (i.i.d.) random samples that are drawn from the joint PMF $p_{X_{\mathcal{K}},Y}$, which is assumed to be given [210]. The k-th encoder observes only the sequence X_k^n, which it would process to generate $J_k = \phi_k(X_k^n)$ per the following (possibly stochastic) mapping [210, eq. (4)]:

$$\phi_k : \mathcal{X}_k^n \rightarrow \mathcal{M}_k^n \tag{34}$$

where $\mathcal{M}_k^n$ denotes an arbitrary set of descriptions [210]. Using $J_{\mathcal{K}} := (J_1, \ldots, J_K)$ as inputs, a (possibly stochastic) decoder $\psi(\cdot)$ processes all the inputs and returns $\hat{Y}^n$ (an estimate of Y^n) as [210, eq. (5)]

$$\psi : \mathcal{M}_1^n \times \ldots \times \mathcal{M}_K^n \rightarrow \hat{\mathcal{Y}}^n. \tag{35}$$

For the mapping in (35), the accuracy of $\hat{Y}^n$ is quantified in terms of *relevance* [210]. Relevance is defined as the information that the descriptions $\phi_1(X_1^n), \ldots, \phi_K(X_K^n)$ *collectively preserve* about Y^n and is given by [210, eq. (6)]

$$\Delta^{(n)}(p_{X_{\mathcal{K}},Y}) := \frac{1}{n} I_{p_{X_{\mathcal{K}},Y}}(Y^n, \hat{Y}^n), \tag{36}$$

where $\hat{Y}^n := \psi(\phi_1(X_1^n), \ldots, \phi_K(X_K^n))$ and the subscript $p_{X_{\mathcal{K}},Y}$ implies that the mutual information is computed w.r.t. the joint distribution $p_{X_{\mathcal{K}},Y}$ [210].

Should the encoder mappings $\{\phi_k\}_{k=1}^K$ be unconstrained, maximizing the RHS of (36) would lead to overfitting [210]. Overfitting can be overcome by using better generalizability, which is usually obtained by constraining the *complexity of the encoders* [210]. To this end, the encoding function $\phi_k(\cdot)$ of encoder $k \in \mathcal{K}$ needs to fulfill [210, eq. (7)]

$$R_k \geq \frac{1}{n} \log |\phi_k(X_k^n)|, \tag{37}$$

where (37) must be satisfied for all $X_k^n \in \mathcal{X}_k^n$ [210]. Meanwhile, optimal performance for distributed learning can be cast as finding the region of all simultaneously achievable *relevance-complexity tuples* [210], as defined below.

Definition 16 ([210, Definition 1]): A tuple $(\Delta, R_1, \ldots, R_K)$ is termed achievable if there exists a training set of size n, encoders ϕ_k for $k \in [K]$, and a decoder ψ such that [210, eqs. (8) and (9)]

$$\Delta \leq \frac{1}{n} I_{p_{X_{\mathcal{K}},Y}}\big(Y^n, \psi(\phi_1(X_1^n), \ldots, \phi_K(X_K^n))\big) \tag{38a}$$

$$R_k \geq \frac{1}{n} \log |\phi_k(X_k^n)|, \quad \forall k \in \mathcal{K}. \tag{38b}$$

The relevance-complexity region $\mathcal{RI}_{\text{DIB}}$ is expressed by the closure of all attainable tuples $(\Delta, R_1, \ldots, R_K)$ [210].

Meanwhile, the region $\mathcal{RI}_{\text{DIB}}$ is characterized by the following theorem.

Theorem 5 ([210, Th. 1]): The relevance-complexity region $\mathcal{RI}_{\text{DIB}}$ of a distributed learning problem with a joint PMF $p_{X_{\mathcal{K}},Y}$ – for which the Markov chain of (33) holds – is expressed by the union of all tuples $(\Delta, R_1, \ldots, R_K) \in \mathbb{R}_+^{K+1}$ fulfilling, for all $\mathcal{S} \subseteq \mathcal{K}$, [210, eq. (14)]:

$$\Delta \leq \sum_{k \in \mathcal{S}} \big[R_k - I(X_k; U_k|Y, T)\big] + I(Y; U_{\mathcal{S}^c}|T), \tag{39}$$

for some of the PMFs $\{p_{U_1|X_1,T}, \ldots, p_{U_K|X_K,T}, p_T\}$ with a joint distribution of the form [210, eq. (15)]:

$$p_T(t)p_Y(y)\prod_{k=1}^{K} p_{X_k|Y}(x_k|y)\prod_{k=1}^{K} p_{U_k|X_k,T}(u_k|x_k, t). \tag{40}$$

Proof. The proof is provided in [210, Sec. 7.1].

Theorem 5 extends the single encoder IB framework to the distributed learning model with K encoders, which is dubbed the DIB problem [210].

The variants of IB, IB, and information theory are the major goal-oriented SemCom frameworks that can inspire many insightful developments in goal-oriented SemCom theories, whose state-of-the-art advancements are presented below.

VI. THEORIES OF GOAL-ORIENTED SEMCOM

In this section, we discuss major developments in goal-oriented SemCom theories. Specifically, we detail below the rate-distortion approach to goal-oriented SemCom [275], [276]; the extended rate-distortion approach to goal-oriented SemCom [277]; and goal-oriented quantization (GOQ) [278], [279]. We begin with the rate-distortion approach to goal-oriented SemCom and, in particular the role of fidelity in goal-oriented SemCom [275].

A. RATE-DISTORTION APPROACH TO GOAL-ORIENTED SEMCOM

The authors of [275] develop a theory that asserts that choosing the type of individual distortion measures (or context-dependent fidelity criteria) per the application/task requirements can considerably affect the semantic source's remote reconstruction. The authors develop their theory by adopting the problem setup proposed by the authors of [280], which is schematized in Fig. 18. The authors of [275] consider a memoryless source represented by the tuple $(\boldsymbol{x}, \boldsymbol{z})$ and has a joint probability density function (PDF) $p(x, z)$ in the product alphabet space $\mathcal{X} \times \mathcal{Z}$. Here, $\boldsymbol{x}$ is the source's semantic or intrinsic information (directly unobservable) and $\boldsymbol{z}$ is the noisy observation of the source at the encoder side [275].

The system model that the authors of [275] adopted in the above-mentioned setup is shown in [275, Fig. 1]. Per [275, Fig. 1] and its accompanying assumption, an information source is a sequence of n i.i.d. RVs $(\boldsymbol{x}^n, \boldsymbol{z}^n)$, and the PDFs $p(x)$ and $p(z|x)$ are assumed to be known [275]. Meanwhile, the encoder (E) and the decoder (D) are defined through the following mappings [275, eq. (1)]:

$$f^E : \mathcal{Z}^n \to \mathcal{W} \tag{41a}$$

$$g_o^D : \mathcal{W} \to \hat{\mathcal{Z}}^n \tag{41b}$$

$$g_s^D : \mathcal{W} \to \hat{\mathcal{X}}^n, \tag{41c}$$

where $\mathcal{W} \in [M]$, g_o^D is the observation decoder, and g_s^D is the semantic information decoder. If one now considers two per-letter distortion measures that are defined by $d_s : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ and $d_o : \mathcal{Z} \times \hat{\mathcal{Z}} \to [0, \infty)$, the corresponding average per-symbol distortions are given by [275, eqs. (2) and (3)]:

$$d_s^n(x^n, \hat{x}^n) := \frac{1}{n}\sum_{i=1}^{n} d_s(x_i, \hat{x}_i) \tag{42a}$$

$$d_o^n(z^n, \hat{z}^n) := \frac{1}{n}\sum_{i=1}^{n} d_o(z_i, \hat{z}_i). \tag{42b}$$

Using (42a) and (42b), the fidelity criteria of the semantic information and observable information are defined as [275]

$$\Delta_s := \mathbb{E}\{d_s^n(x^n, \hat{x}^n)\} \text{ and } \Delta_o := \mathbb{E}\{d_o^n(z^n, \hat{z}^n)\}, \tag{43}$$

respectively. Using (42a)-(43), we state the following definition regarding the achievable rates and the infimum of all achievable rates.

Definition 17 ([275, Definition 1]): For two distortion levels $D_o, D_s \geq 0$, R is said to be (D_o, D_s)–achievable w.r.t. an arbitrary $\epsilon > 0$, there exists – for a very large n – a semantic-aware lossy source code $(n, M, \Delta_o, \Delta_s)$ with $M \leq 2^{n(R+\epsilon)}$ given that $\Delta_o \leq D_o + \epsilon$ and $\Delta_s \leq D_s + \epsilon$. Furthermore, considering that sequences of distortion functions $\{(d_o^n, d_s^n) : n = 1, 2, \ldots\}$ are given, then [275, eq. (5)]

$$R(D_o, D_s) := \inf\{R : (R, D_o, D_s) \text{ is achievable}\}. \tag{44}$$

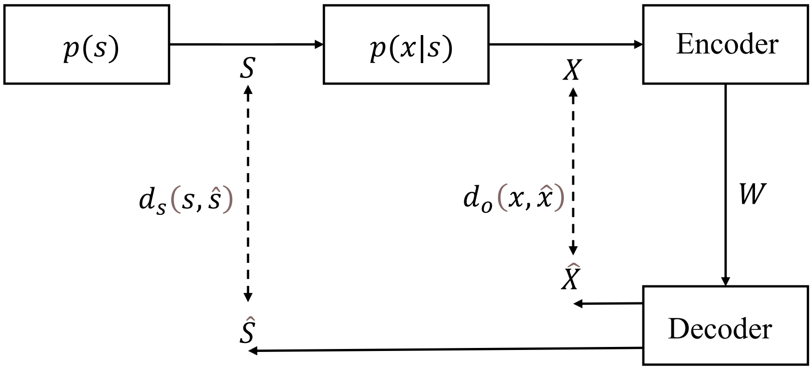


FIGURE 18. An illustrating schematic of a semantic source and its loss compression – redrawn from [280, Fig. 1].

Per Definition 17, the information-theoretic characterization of (44) is captured by the following lemma.

Lemma 1 ([275, Lemma 1]): For a given $p(x)$ and $p(z|x)$, the semantic rate distortion function of the system model in [275, Fig. 1] can be expressed as [275, eq. (6)]

$$R(D_s, D_o) = \inf_{q(\hat{z},\hat{x}|z)} I(z;\hat{z},\hat{x}) \tag{45a}$$

$$\text{s.t.} \quad \mathbb{E}\{\hat{d}_s(z,\hat{x})\} \leq D_s \tag{45b}$$

$$\mathbb{E}\{d_o(z,\hat{z})\} \leq D_o, \tag{45c}$$

where $\hat{d}_s(z,\hat{x}) := \sum_{x\in\mathcal{X}} p(x|z)d_s(x,\hat{x})$, $D_s \in [0,\infty]$, $D_o \in [0,\infty]$, and [275, eq. (7)]

$$I(z;\hat{z},\hat{x}) := \mathbb{E}\left\{\log\left(\frac{q(\hat{z},\hat{x}|z)}{\nu(\hat{z},\hat{x})}\right)\right\}. \tag{46}$$

The constrained optimization problem in (45a)-(45c) can be written as an unconstrained optimization problem through the *Lagrange duality theorem* as follows [275, eq. (15)]:

$$R(D_s, D_o) = \max_{s_1,s_2\leq 0} \min_{q(\hat{z},\hat{x}|z)\geq 0, \sum_{\hat{z},z} q(\hat{z},\hat{x}|z)=1} \Big\{I(z;\hat{z},\hat{x}) - s_1\big(\mathbb{E}\{\hat{d}_s(z,\hat{x})\} - D_s\big) - s_2\big(\mathbb{E}\{d_o(z,\hat{z})\} - D_o\big)\Big\}, \tag{47}$$

where $s_1, s_2 \leq 0$ are the Lagrange multipliers. The authors of [275] solve (47) and state the following main result.

Theorem 6 ([275, Th. 1]): Given that $p(x)$ and $p(z|x)$ are known, the underneath parametric solutions follow for the optimization problem in (45a)-(45c):

- If $s_1, s_2 < 0$, the implicit optimal form of the minimizer that attains the minimum is given by [275, eq. (16)]. In addition, the optimal parametric solution when $R(D_s^*, D_o^*) > 0$ is expressed by [275, eq. (17)].
- If $s_1 < 0, s_2 = 0$, and $R(D_s^*, D_o^*) > 0$, $R(D_s^*, D_o^*)$ is given by [275, eq. (20)].
- If $s_1 = 0, s_2 < 0$, and $R(D_s^*, D_o^*) > 0$, $R(D_s^*, D_o^*)$ is characterized by [275, eq. (21)].
- If $s_1 = s_2 = 0$, $R(D_s^*, D_o^*) = 0$.

Proof. The proof is given in [275, Appendix A].

Theorem 6 is useful for deriving analytical expressions of the constrained optimization problem in (45a)-(45c) and constructing generalizations of the *Blahut–Arimoto algorithm* (BA algorithm) [269].

We now move on to discuss an extended rate-distortion approach to goal-oriented SemCom [277].

B. EXTENDED RATE-DISTORTION APPROACH TO GOAL-ORIENTED SEMCOM

The authors of [277] put forward a JSCC-based goal-oriented SemCom system that incorporates a semantic reconstruction scheme while focusing on predicting the precision and generalizability of multiple goals/tasks. This goal-oriented SemCom system is composed of a JSCC encoder, a quantizer, a wireless channel, a JSCC decoder, and a network of AI tasks at the receiver [277, Fig. 2]. When the system is fed input X, which denotes an RV pertaining to the source image space, let an RV Y be the desired output of an AI task. As can be seen in [277, Fig. 2], the JSCC encoder maps the input to semantic representations that are subsequently quantized by the quantizer to minimize the transmission cost. The quantized symbols Z are then transmitted over a wireless channel to the receiver [277]. At the receiver, the JSCC decoder maps the noisy received symbols to the reconstructed image $\hat{X}$. Eventually, the AI task network uses $\hat{X}$ as an input and produces its prediction $\hat{Y}$. This overall goal-oriented SemCom scheme is formulated as an extended rate-distortion problem [277], and its analytical characterization is presented below.

To ensure that the reconstructed images can perform the given AI task properly, IB distortion [208] must be minimized. To this end, the IB distortion between x and $\hat{x}$ amounts to the KL divergence $D_{KL}(p(y|x)||p(y|\hat{x}))$, which is given by [277, eq. (1)]

$$d_{IB}(x,\hat{x}) := \sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{p(y|\hat{x})}, \tag{48}$$

where $\mathcal{Y}$ is the alphabet of Y. For (48), $D_{IB}(X,\hat{X}) := \mathbb{E}\big\{d_{IB}(x,\hat{x})\big\}$ is the conditional mutual information $I(X;Y|\hat{X})$

[271] and defined as [277, eq. (2)]

$$D_{IB}(X,\hat{X}) = \sum_{x\in\mathcal{X}}\sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{y\in\mathcal{Y}} p(x,\hat{x})p(y|x)\log\frac{p(y|x)}{p(y|\hat{x})}, \tag{49}$$

where $\mathcal{X}$, $\hat{\mathcal{X}}$, and $\mathcal{Y}$ are the alphabet of X, $\hat{X}$, and Y, respectively. The definition in (49) then leads to the following theorem.

Theorem 7 ([277, Th. 1]): $D_{IB}(X,\hat{X})$ as defined in (49) can also be expressed as [277, eq. (3)]

$$D_{IB}(X,\hat{X}) = I(X;Y) - I(\hat{X};Y). \tag{50}$$

Proof. The proof is given in [277, Appendix A].

The relation in (50) intuitively illustrates the reduction of useful information [277]. To improve the generalizability among different AI tasks, one must also minimize the reconstruction distortion $D_{RD}(X,\hat{X})$ that is equated as [277, eq. (4)]

$$D_{RD}(X,\hat{X}) := \sum_{x\in\mathcal{X}}\sum_{\hat{x}\in\hat{\mathcal{X}}} p(x,\hat{x})d_{RD}(x,\hat{x}), \tag{51}$$

where $d_{RD}(x,\hat{x}) = (x-\hat{x})^2$ [277, eq. (5)]. Meanwhile, the authors of [277] take into account the natural tradeoff between $D_{IB}(X,\hat{X})$ and $D_{RD}(X,\hat{X})$, and define the semantic distortion measurement $D_S(X,\hat{X})$ as [277, eq. (6)]

$$D_S(X,\hat{X}) := D_{RD}(X,\hat{X}) + \beta D_{IB}(X,\hat{X}), \tag{52}$$

where β controls the tradeoff between the AI task's prediction accuracy and the goal-oriented SemCom system's generalizability [277]. Using (52), the goal-oriented SemCom system proposed in [277] can be formulated as an extended rate-distortion optimization problem given by [277, eq. (10)]

$$\min_{p(\hat{x}|x)} \quad D_{RD}(X,\hat{X}) + \beta D_{IB}(X,\hat{X}) \tag{53a}$$

$$\text{s.t.} \quad I(X;\hat{X}) \leq I_C \tag{53b}$$

$$\sum_{\hat{x}} p(\hat{x}|x) = 1, \tag{53c}$$

where the constraints in (53b) and (53c) correspond to the maximum channel capacity I_C and the normalization constraint of the conditional PMF $p(\hat{x}|x)$, respectively [277]. Substituting (50) into (53a) and discarding $I(X;Y)$ – since it is constant for a given dataset – leads to the following optimization problem:

$$\min_{p(\hat{x}|x)} \quad D_{RD}(X,\hat{X}) - \beta I(\hat{X};Y) \tag{54a}$$

$$\text{s.t.} \quad I(X;\hat{X}) \leq I_C \tag{54b}$$

$$\sum_{\hat{x}} p(\hat{x}|x) = 1. \tag{54c}$$

The fact that the authors of [277] solve this optimization problem using the Lagrange multiplier technique leads to the following theorem.

Theorem 8 ([277, Th. 2]): The optimal mapping from the source images X to the semantically-reconstructed images $\hat{X}$ must satisfy [277, eqs. (12)–(15)]:

$$p(\hat{x}|x) = \frac{p(\hat{x})e^{-\lambda^{-1}d_S(x,\hat{x})}}{\mu(x)} \tag{55a}$$

$$p(\hat{x}) = \sum_{x\in\mathcal{X}} p(x)p(\hat{x}|x) \tag{55b}$$

$$p(y|\hat{x}) = \sum_{x\in\mathcal{X}} p(y|x)p(x|\hat{x}), \tag{55c}$$

where

$$\mu(x) = \sum_{\hat{x}\in\hat{\mathcal{X}}} p(\hat{x})e^{-\lambda^{-1}d_S(x,\hat{x})} \tag{56a}$$

$$d_S(x,\hat{x}) = d_{RD}(x,\hat{x}) + \beta d_{IB}(x,\hat{x}). \tag{56b}$$

Proof. The proof is provided in [277, Appendix B].

The optimal distributions $p(\hat{x}|x)$, $p(\hat{x})$, and $p(y|\hat{x})$ can be obtained [277] using the BA algorithm [269].

Meanwhile, we now continue with our discussion of GOQ [278], [279].

C. GOAL-ORIENTED QUANTIZATION

GOQ is quite useful for many applications, including controlled networks that are built on a communication network, wireless resource allocation, and 6G systems [278], [279]. In this vein, a general GOQ framework wherein the goal/task of a receiver is modeled by a generic optimization problem that comprises both decision variables and parameters is illustrated in [278, Fig. 1]. More specifically, the goal is modeled as a minimization problem of a general goal function $f(\boldsymbol{x};\boldsymbol{g})$ for $\boldsymbol{x}$ (with dimension d) being the decision that has to be made from a quantized version of the parameters $\boldsymbol{g}$ (with dimension p) [278]. In view of this problem, we state the following two definitions.

Definition 18 ([278, Definition II.1]): Suppose $M,d \in \mathbb{N}_{\geq 1}$ and $\mathcal{G} \in \mathbb{R}^d$. An M–quantizer $\mathcal{Q}_M$ is completely decided by a piecewise constant function $\mathcal{Q}_M : \mathcal{G} \to \mathcal{G}$. This mapping is defined as $\mathcal{Q}_M(g) = z_m$ for all $z_m \in \mathcal{G}_m$ given that $m \in [M]$; $\mathcal{G}_1,\ldots,\mathcal{G}_M$ are the quantization regions that define a partition of $\mathcal{G}$; and $z_1,\ldots,z_M$ are the region representatives.

Definition 19 ([278, Definition II.2]): Suppose $p \in \mathbb{N}_{\geq 1}$ and g is a fixed parameter. Let $\chi(g)$ be a decision function that provides the minimum points for the goal function $f(\boldsymbol{x};g)$, whose decision variable is $\boldsymbol{x} \in \mathbb{R}^p$ [278, eq. (1)]:

$$\chi(g) \in \arg\min_{\boldsymbol{x}} f(\boldsymbol{x};g). \tag{57}$$

The optimality loss induced by quantization is equated as [278, eq. (2)]:

$$L(\mathcal{Q};f) := \alpha_f \int_{g\in\mathcal{G}} \big[f(\chi(\mathcal{Q}(g));g) - f(\chi(g);g)\big]\phi(g)dg, \tag{58}$$

where $\phi(\cdot)$ is the PDF of g and $\alpha_f > 0$ denotes a scaling factor that is independent of $\mathcal{Q}$.

From Definition 19, the following remarks follow.

Remark 1: The conventional quantization approach can be derived from the GOQ approach by observing that the second term of $L(Q;f)$ – as defined in (58) – is independent of Q and specifying f as $f(\boldsymbol{x};\boldsymbol{g}) = \|\boldsymbol{x}-\boldsymbol{g}\|^2$ [278].

Remark 2: Unlike the conventional quantization approach that aims to provide a version of g that resembles g, what matters in the GOQ approach is the quality of the end decision made [278].

Remark 3: The design of a GOQ quantizer constitutes a major difference w.r.t. the conventional quantization approach and thus hinges on the mathematical properties of f and the underlying decision function $\chi(\cdot)$ [278].

When it comes to Remark 3, quantifying the relationship between the nature of f and the quantization performance is a challenging problem [278]. Meanwhile, for a scalar GOQ such that $d = p = 1$ and $\rho(\cdot)$ being a density function, the number of quantization intervals over $[a,b]$ can be approximated by $M\int_a^b \rho(g)dg$ [278]. Accordingly, the problem of finding a GOQ in the high-resolution regime boils down to finding the density function that minimizes the optimality loss that is denoted by $L(\rho;f)$ [278]. This leads to the following proposition.

Proposition 1 ([278, Proposition III.1]): Suppose f is a fixed goal function that is assumed to be κ times differentiable and χ differentiable with [278, eq. (4)]

$$\kappa = \min\left\{i \in \mathbb{N} : \forall g, \left.\frac{\partial^i f(x;g)}{\partial x^i}\right|_{x=\chi(g)} \neq \text{a.s.}\right\}. \tag{59}$$

In the high resolution regime, the optimality loss $L(\rho;f)$ is minimized by employing the underneath quantization interval density function [278, eq. (5)]:

$$\rho^\star(g) = C\left[\left(\frac{d\chi(g)}{dg}\right)^{\kappa}\frac{\partial^\kappa f(\chi(g);g)}{\partial x^\kappa}\phi(g)\right]^{\frac{1}{\kappa+1}}, \tag{60}$$

where $\frac{1}{C} = \int_G \left[\left(\frac{d\chi(t)}{dt}\right)^{\kappa}\frac{\partial^\kappa f(\chi(t);t)}{\partial x^\kappa}\phi(t)\right]^{\frac{1}{\kappa+1}} dt.$

Proof. The proof is provided in [278, Appendix A].

On the other hand, when $d, p \in \mathbb{N}_{\geq 1}$, the goal-oriented quantization problem becomes a vector GOQ problem [278], which the following proposition is derived for.

Proposition 2 ([278, Proposition IV.1]): Assume $d, p \in \mathbb{N}_{\geq 1}$; $\kappa = 1$; and f and χ are twice differentiable. Let $\boldsymbol{H}_f(\boldsymbol{x};\boldsymbol{g})$ and $\boldsymbol{J}_\chi(\boldsymbol{g})$ be the Hessian matrix of f and the Jacobian matrix of f evaluated for an optimal decision $\chi(\boldsymbol{g})$, respectively. In the regime of large M, $L(Q;f)$ – which is the optimality loss function defined in (58) – is approximated as [278, eq. (9)]:

$$L(Q;f) = \underbrace{\alpha_f \sum_{m=1}^{M}\int_{\mathcal{G}_m}(\boldsymbol{g}-z_m)^T\boldsymbol{A}_{f,\chi}(\boldsymbol{g})(\boldsymbol{g}-z_m)\phi(\boldsymbol{g})d\boldsymbol{g}}_{=\hat{L}_M(Q;f)} + \mathcal{O}(M^{-2/p}), \tag{61}$$

where $\boldsymbol{A}_{f,\chi}(\boldsymbol{g}) = \boldsymbol{J}_\chi^T(\boldsymbol{g})\boldsymbol{H}_f(\chi(\boldsymbol{g});\boldsymbol{g})\boldsymbol{J}_\chi(\boldsymbol{g})$. Moreover, $\hat{L}_M(Q;f)$ – as expressed in (61) – can be bounded as $L_M^{\min}(Q;f) \leq \hat{L}_M(Q;f) \leq L_M^{\max}(Q;f)$, where $L_M^{\min}(Q;f)$ and $L_M^{\max}(Q;f)$ are given in [278, eq. (10)] and [278, eq. (11)], respectively.

Proof. The proof is provided in [278, Appendix B].

Apart from the above-discussed goal-oriented SemCom theories, there have also been other theoretical developments such as the *theory of goal-oriented communication* [281] and *universal SemCom II* [158]. This leads us to our in-depth discussion of the fundamental and major challenges of goal-oriented SemCom. It is worth noting, however, that the above-discussed goal-oriented SemCom theories have their corresponding limitations and are hence not the most rigorous and complete of theories (though they are interesting!). This is attributed to the numerous fundamental and major challenges of goal-oriented SemCom, which are detailed below along with its future research directions.

VII. CHALLENGES AND FUTURE RESEARCH DIRECTIONS OF GOAL-ORIENTED SEMCOM

A. FUNDAMENTAL AND MAJOR CHALLENGES OF GOAL-ORIENTED SEMCOM

When it comes to realizing high-fidelity goal-oriented SemCom for 6G and beyond, the research field of goal-oriented SemCom is fraught with fundamental and major challenges in the theoretical, algorithmic, and realization/implementation-related research frontiers. These challenges are discussed in detail below, beginning with the challenges in the development of fundamental goal-oriented SemCom theories.

1) CHALLENGES IN THE DEVELOPMENT OF FUNDAMENTAL GOAL-ORIENTED SEMCOM THEORIES

We detail below (in no specific order) the fundamental and major challenges related to – but not limited to – the development of fundamental goal-oriented SemCom theories.

- *Lack of a Commonly Accepted Definition of Semantics (Semantic Information)*: There exist a number of definitions for semantic information:
  - The authors of [282] and [283] argue that the fundamental notion of semantic information depends on the information ecosystem, which is a complete process of *information-knowledge-intelligence conversion*.
  - In putting forth an information model dubbed the *evolutionary energetic information model*, the author of [284] reasons that semantic information is an exclusive feature of biological evolution.
  - Semantic information via multiple definitions of *semantic entropy* – i.e., *semantic entropy of a sentence* [285]; *knowledge entropy* [286]; semantic entropy grounded on a language comprehension model [287]; an information-theoretic method for measuring semantic entropy in translation

tasks [288]; a fuzzy set theory-based definition of semantic entropy [289]; and so forth.

Despite the variety of itemized definitions [61, Sec. II], there is no commonly agreed upon definition for semantics (semantic information). This is a fundamental challenge that can hinder the advancement of goal-oriented SemCom theory, algorithm, and realization.

- *Fundamental Performance Analysis of Goal-Oriented SemCom*: According to Fig. 3, the fundamental performance analysis of goal-oriented SemCom has to account for the likelihood of successful action execution using the semantic decoding / semantic inference module's outputs, which are steered by the destination KB whose changing content of knowledge has to match – in real-time – with the knowledge of the source KB, which in turn guides the semantic representation / semantic filtering / semantic encoding module of the semantic transmitter. Consequently, quantifying the fundamental non-asymptotic performance of a oal-oriented SemCom system is incredibly challenging for the following reasons [67]:
  - The lack of a commonly agreed-upon definition of semantics (semantic information) [188, Ch. 10, p. 125].
  - There is no universal technique for a rigorous semantic representation [61, Sec. II].
  - The lack of a comprehensive mathematical foundation for goal-oriented SemCom [290, Sec. IV].
  - The inevitability of knowledge mismatch between the source KB and destination KB, which is very difficult to quantify in real-time.

  Moreover, since a system's goal may not be explicitly represented by a utility function, it can be fundamentally challenging to rigorously analyze a goal-oriented SemCom system's performance.
- *Performance Analysis of DL-enabled Goal-Oriented SemCom Systems*: DL-based goal-oriented SemCom systems such as cooperative goal-oriented SemCom [175] rely on a DL-based semantic encoder, a joint DL-based source-channel coding, and a DL-based semantic decoder. The rigorous non-asymptotic performance analysis of DL-based goal-oriented SemCom systems is thus hindered by the *fundamental lack of interpretability/explainability* inherent in DL models [291], [292], [293], concerning the lack of interpretability/explainability in their optimization, generalization, and approximation [294], [295], [296].
- *Fundamental Limits of Goal-Oriented SemCom Systems*: The fundamental limits of goal-oriented SemCom depend on not only the type of DL-based semantic encoder and semantic decoder used, but also the type of goal, and hence the goal function. The goal function can hardly be detailed enough to capture all aspects of a goal, and DL-based goal-oriented SemCom techniques suffer from a fundamental lack of interpretability (the same as DL-based SemCom schemes).
- *Semantic Compressed Sensing and Optimal Sampling Theory*: In stark contrast to the state-of-the-art techniques that pursue a "sample-then-compress" structure, *semantic compressed sensing* is a computationally lighter scheme that gathers only the minimum volume of data needed to reconstruct the signal of interest at the desired resolution, as determined by the application requesting the data [195]. It carries out certain signal processing operations directly in the "compressed domain" without complete signal reconstruction [195]. This calls for tackling the formidable challenge of developing an *optimal sampling theory* that unifies signal sparsity and aging/semantics for real-time prediction/reconstruction under communication and delay constraints [195].

We now carry on with fundamental and major challenges in the development of fundamental goal-oriented SemCom algorithms.

2) CHALLENGES IN THE DEVELOPMENT OF FUNDAMENTAL GOAL-ORIENTED SEMCOM ALGORITHMS

We detail below (in no specific order) the fundamental and major challenges related to - but not limited to - the development of fundamental goal-oriented SemCom algorithms.

- *Inevitability of Semantic Mismatch*: The source KB and the destination KB can be quite different because they observe different worlds with unequal abilities to understand things [73]. Consequently, semantic mismatch is unavoidable to the extent that it can fundamentally constrain the performance of wireless systems that are based on goal-oriented SemCom.
- *Lack of Unified Semantic Performance Assessment Metrics*: Despite the numerous metrics that have been proposed for goal-oriented SemCom [297], there is a lack of unified/universal performance assessment metrics for goal-oriented SemCom [43]. When it comes to unified metrics, the major challenge is to establish concrete metrics that can capture source and network dynamics, as well as any potentially non-trivial interdependencies among information attributes [83].
- *Lack of Interpretability in DL-Based Goal-Oriented SemCom*: There is a fundamental lack of interpretability in DL-based goal-oriented SemCom algorithms due to the fundamental lack of interpretability/explainability that is inherent in trained DL models [291], [292], [295], [296]. This is a foundational challenge, should one aims for developing DL-enabled interpretable goal-oriented SemCom algorithms.
- *Optimal Semantic-Aware Joint Sampling, Transmission, and Reconstruction of Multidimensional Signals*: In a number of conventional communication systems,

transmission is optimized on the basis of quality of service (QoS) metrics – e.g., delay, rate, and timeliness – while ignoring source variations, the fact that samples may be received on time but contain no useful information; or the fact that samples can even be misleading about the system's true state [83]. This scenario highlights the implicit structural links that exist between sampling and communication, which are generally inseparable in SemCom and goal-oriented SemCom [83]. For reliable goal-oriented SemCom that enables timely decision-making and satisfies the stringent requirements of real-time NCSs, the formidable challenge is therefore to develop a theory for optimal semantic-aware joint active sampling, transmission, and reconstruction of multi-dimensional signals, especially under stringent timing constraints [83].

- *Resource Allocation for Goal-Oriented SemCom*: From the vantage point of optimal resource allocation, goal-oriented SemCom systems face many fundamental challenges, some of which have led to the following major research problems [75]:
  - *How can a generic resource allocation problem be optimized for different goal-oriented SemCom systems?*
  - *How can a resource allocation policy be optimized while maximizing goal-oriented SemCom's efficiency?*
- *Goal-Oriented Resource Orchestration*: In emerging cyber-physical and autonomous networked systems, semantic-aware real-time data networking requires effective scheduling and resource allocation policies for gathering (often correlated) multi-source multi-modal information [83]. The objectives in the networked applications could be achieved by using an alternative set of multi-quality data [83]. These goal-oriented resource orchestration problems fall into the realm of real-time scheduling with multiple choices [83]. It is therefore challenging to devise online algorithms that can select which piece of information – from where and when – to gather and transmit under communication and processing constraints [83].
- *Multi-Objective Stochastic Optimization*: When it comes to goal-oriented end-user-perceived utilities that estimate the relative degree of priority of different information attributes, semantic-aware data gathering and prioritization require multi-criteria optimization [83]. In view of this optimization and overcoming its challenges, multi-objective stochastic optimization based on the cumulative prospect theory – which incorporates semantic information via risk-sensitive measures and multi-attribute entropy-based utility functions – holds promise [83].

We now proceed to discuss the fundamental and major challenges in the realization of goal-oriented SemCom.

3) CHALLENGES IN THE REALIZATION OF GOAL-ORIENTED SEMCOM

We henceforth discuss (in no specific order) the fundamental and major challenges related to – but not limited to – the realization of goal-oriented SemCom.

- *Real-Time Requirement*: Several major use cases of goal-oriented SemCom, such as autonomous transportation, telehealth, smart factories, and NCSs, have real-time requirements for goal-oriented communication/control. However, incorporating semantic reasoning into the goal-oriented SemCom use cases mentioned incurs extra delay in goal-oriented SemCom's overall transceivers [77]. Satisfying the ultra-low end-to-end latency requirements (i.e., real-time requirements) of 6G and beyond is therefore a major realization challenge for goal-oriented SemCom.
- *Scalability*: As is the case for SemCom, the realization of goal-oriented SemCom is hampered by several scalability challenges, such as:
  - The lack of a general semantic-level framework for distinct types of sources.
  - Sharing, updating, and maintaining KBs at the source and destination definitely necessitate additional storage costs and algorithm design [77].
  - Realizing goal-oriented SemCom involves significant computational as well as storage costs.
- *Knowledge Evolution Tracking*: Many existing goal-oriented SemCom techniques rely on the dynamic sharing of knowledge between the source KB and the destination KB. To this end, modeling and keeping track of each piece of knowledge is fundamentally important for improving the efficiency and reliability of goal-oriented SemCom. Nonetheless, the basic neuroscientific understanding of knowledge, knowledge evolution, and knowledge tracking are very difficult fundamental problems.
- *Compatibility with Existing Communication Infrastructure*: Since bit communication (BitCom) systems and services will still be in use when goal-oriented SemCom systems and services are rolled out in 6G networks and systems, any implementation of goal-oriented SemCom should ensure that futuristic goal-oriented SemCom systems are compatible with the existing communication infrastructure. To this end, extensive link-level simulations must be performed to verify the realistic end-to-end performance of goal-oriented SemCom, under the presence of BitCom transmissions.
- *Efficient Knowledge Sharing in Multi-User MIMO Goal-Oriented SemCom Systems*: A multi-user MIMO goal-oriented SemCom system such as cooperative goal-oriented SemCom [175] – which is schematized in Fig. 11 – needs knowledge to be shared between the receiver with multiple antennas and a number of goal-oriented SemCom users that are equipped

with either a single antenna or multiple antennas. However, achieving efficient global knowledge sharing in multi-user MIMO goal-oriented SemCom systems is challenging.

Because challenges are always opportunities, some of the above-detailed fundamental and major challenges of goal-oriented SemCom are also big opportunities for novel future directions for goal-oriented SemCom, as discussed below.

B. FUTURE DIRECTIONS OF GOAL-ORIENTED SEMCOM

In light of the fundamental and major challenges of goal-oriented SemCom that are highlighted in Section VII-A, the developments in theories of goal-oriented SemCom that are discussed in Section VI, the major trends and use cases of goal-oriented SemCom that are detailed in Section IV, and the many proposals of state-of-the-art goal-oriented SemCom algorithms that are surveyed in Section III, we offer some novel future directions for goal-oriented SemCom theories, algorithms, and realization. We begin with some novel future directions for goal-oriented SemCom theories.

1) FUTURE DIRECTIONS FOR GOAL-ORIENTED SEMCOM THEORIES

We highlight (in no particular order) some novel future directions related to – but not limited to – goal-oriented SemCom theories.

- *A Fundamental Theory and the Fundamental Limits of Actionable Intelligence*: Actionable intelligence is on-time and accurate intelligence that would help decision-makers make an optimal/well-informed decision [298]. Representing decision in the context of goal-oriented SemCom, the reconstructed signals of a communicating smart device can alter the recipients' states and initiate specific actions at the receivers [83]. The limits of actionable intelligence must be well-understood before deploying any goal-oriented SemCom system. To this end, a fundamental theory and the fundamental limits of actionable intelligence – in the context of DL, big data, or a combination thereof – are critical future research directions for goal-oriented SemCom.
- *A Fundamental Theory of Optimal Semantic-Aware Joint Active Sampling, Transmission, and Reconstruction of Multi-Dimensional Signals*: A theory of optimal semantic-aware joint active sampling, transmission, and reconstruction of multi-dimensional signals – especially, under stringent timing constraints – is needed to enable timely decision-making and efficiently meet the requirements of real-time networked applications [83].

We now proceed to highlight some novel future directions for goal-oriented SemCom algorithms.

2) FUTURE DIRECTIONS FOR GOAL-ORIENTED SEMCOM ALGORITHMS

We point out (in no specific order) the promising future directions related to – but not limited to – goal-oriented SemCom algorithms.

- *Semantic-Aware Networking*: In semantic-aware goal-oriented networks, the major operations include local goal-oriented information acquisition, representation, and semantic value inference; data prioritization; in-network processing (e.g., fusion and compression); semantic reception; and semantic reconstruction [83]. These operations will require optimal or nearly-optimal algorithms for semantic filtering, semantic preprocessing, semantic reception, and semantic control [83].
- *Goal-Oriented SemCom with Time-Evolving Goals*: Although most state-of-the-art goal-oriented SemCom works consider fixed goals, it is often the case that one task is followed by one or more other tasks in different systems that include smart devices [85]. This enforces the design constraint that a new task needs to be executed seamlessly once the previous task has ended [85]. Nevertheless, retraining from scratch for every goal not only takes time but also wastes resources [85]. Consequently, a unified goal-oriented SemCom framework that takes into account multiple – often causally related – goals while maximizing the expected goal accomplishment [85] is a research direction worth pursuing.
- *Goal-Oriented Coding and Control*: Source coding or JSCC models could be implemented to characterize goal-oriented compression and its performance limits [85]. Whenever a goal relies on not only the state, but also the decision made, in the current time slot as well as previous time slots, formulating dynamic system models can lead to a promising solution [85]. For this scenario, there are two possible ways to design optimal goal-oriented coding and control [85], which are worth a through investigation:
  - Resorting to differential equations to explore the evolution of the transmitted messages and the goal.
  - Revisiting the sampling process by tailoring the sampling problem to a general utility function.
- *Multi-Modal Goal-Oriented SemCom for the Metaverse*: Most existing goal-oriented SemCom techniques revolve around semantic extraction, semantic encoding, and semantic decoding for a single task. The Metaverse, however, requires multi-modal service models that include multiple types of immediate interactions, such as audio, image, video, and haptic services [32]. This calls for multi-modal goal-oriented SemCom techniques that address the following research themes [32]:
  - The design and implementation of multi-modal goal-oriented SemCom models that can provide multi-sensory multimedia services in the Metaverse.
  - Efficiently extracting semantic information from the data generated by the Metaverse users.
  - Optimal resource allocation in the edge network that enables the training and application of multi-modal goal-oriented SemCom models.

It is worth pursuing the itemized research themes toward rigorous multi-modal goal-oriented SemCom technique.

We now move on to some crucial future directions for goal-oriented SemCom realization.

3) FUTURE DIRECTIONS FOR GOAL-ORIENTED SEMCOM REALIZATION

In what follows, we point out (in no particular order) some useful future directions related to – but not limited to – goal-oriented SemCom realization.

- *The Coexistence of BitCom and Goal-Oriented SemCom Users*: Since BitCom service and infrastructure will still be in use when goal-oriented SemCom is implemented in 6G and beyond, the coexistence of BitCom users and goal-oriented SemCom users must be investigated through the lens of not only measurements, but also theory. Regarding theory, the coexistence of BitCom users and goal-oriented SemCom users should be studied in detail from the vantage points of optimal resource allocation and interference mitigation.
- *The Impact of Inconsistent KBs at the Source and Destination*: Even though most state-of-the-art goal-oriented SemCom proposals resort to the assumption that knowledge is shared in real time to consider consistent KBs at the source and destination, the source KB and the destination KB are fundamentally inconsistent [73]. Therefore, how to design and realize novel (multi-user) goal-oriented SemCom systems with inconsistent KBs are an open issue in goal-oriented SemCom design and realization.

At last, we move on to our concluding summary and research outlook.

VIII. CONCLUDING SUMMARY AND RESEARCH OUTLOOK

Inspired by many existing heterogeneous 6G driving applications, trends, and use cases, various researchers in academia, industry, and national laboratories have disseminated several 6G proposals. In spite of the many 6G proposals, materializing 6G – as currently envisioned – is fraught with many fundamental IMT challenges. To mitigate some of these challenges, SemCom and goal-oriented SemCom have emerged as promising 6G technology enablers. By taking a pragmatic approach to SemCom, goal-oriented SemCom focuses on employing only semantically-relevant information for successful task execution while minimizing bandwidth consumption, power usage, and transmission delay. This asserts the criticality of goal-oriented SemCom for 6G. On the other hand, 6G is also essential for the materialization of major goal-oriented SemCom use cases. These paradigms of *6G for goal-oriented SemCom* and *goal-oriented SemCom for 6G* call for a tighter integration of 6G and goal-oriented SemCom. To facilitate this purpose, this survey article elaborated the essence of goal-oriented SemCom and its state-of-the-art research landscape. It then documented the major state-of-the-art trends, use cases, and frameworks of goal-oriented SemCom; revealed the fundamental and major challenges of goal-oriented SemCom; and provided promising future research directions for goal-oriented SemCom theories, algorithms, and realization.

This survey article also discussed the fundamental challenges of 6G and provided informative tables of 6G driving applications (along with their corresponding industry verticals), 6G algorithm/protocol-level enablers (along with their respective KPI impacts), 6G infrastructure-level enablers (along with their respective KPI impacts), 6G spectrum-level enablers (along with their respective KPI impacts), and challenges and open problems for the 6G security, trust, and privacy. By exposing the 6G fundamental challenges; revealing the fundamental and major challenges of goal-oriented SemCom; and offering novel future research directions for goal-oriented SemCom theories, algorithms, and realization, this comprehensive survey duly stimulates many lines of research on goal-oriented SemCom theories, algorithms, and realization.

APPENDIX. – LIST OF ABBREVIATIONS & ACRONYMS

Abbreviation	Definition
3CLS	Communications, computing, control, localization, and sensing.
(1/2/3)D	(One/two/three)-dimensional.
5G	Fifth-generation.
5GNR	5G new radio.
6G	Sixth-generation.
6GRAN	6G radio access network.
AE	Autoencoder.
AGI	Artificial general intelligence.
AI	Artificial intelligence.
AIaaS	AI-as-a-service.
AI4Net	AI for network.
AIGC	AI-generated content.
AoI	Age of information.
AR	Augmented reality.
ASC	Adaptable semantic compression.
AVs	Autonomous vehicles.
BCI	Brain-computer interaction/interface.
BERT	Bidirectional encoder representations from Transformers.
BitCom	Bit communication.
CCs	Color codes.
CCCA	Cache-computing coordination algorithm.
CF	Computing force.
CFN	Computing force network.
CMOS	Complementary metal oxide semiconductor.
CNN(s)	Convolutional neural network(s).
CP-DQN	Content popularity-based deep Q-network.
CSI	Channel state information.
De-Fi	Decentralized finance.
DL	Deep learning.
DNN(s)	Deep neural network(s.

DQN	Deep Q-network.
DT-JSCC	Discrete task-oriented joint source-channel coding.
DIB	Distributed information bottleneck.
eMBB	Enhanced mobile broadband.
ESC	Emergent semantic communication.
FC	Fusion center.
FP-error	False positive error.
FoV	Field-of-view.
GFlowNets	Generative flow networks.
GIB	Graph information bottleneck.
GOQ	Goal-oriented quantization.
H2H	Human-to-human.
H2M	Human-to-machine.
IB	Information bottleneck.
ID	Identification via channels.
i.i.d.	Independent and identically distributed.
IMT	Interdisciplinary, multidisciplinary, and transdisciplinary.
IIoT	Industrial IoT.
IoE	Internet of everything.
IoT	Internet of things.
IoV	Internet of vehicles.
ISCC	Integrated sensing, computation, and communication.
ISGC	Integrated SemCom and AIGC.
JND	Just noticeable difference.
JSC	Joint source-channel.
JSCC	Joint source-channel coding.
KB	Knowledge base.
KL	Kullback-Leibler.
KPI	Key performance indicator.
LED	Light emitting diode.
L-MMSE	Linear minimum-mean-squared error.
M2M	Machine-to-machine.
MA-POMDP	Multi-agent partially observable Markov decision process.
MARL	Multi-agent reinforcement learning.
MEC	Mobile edge computing.
MIMO	Multiple-input multiple-output.
ML	Machine learning.
mMTC	Massive machine type communications.
mmWave	millimeter wave.
MR	Mixed reality.
MSE	Mean-squared error.
Net4AI	Network for AI.
NFT	Non-fungible token.
NCSs	Networked control systems.
NeSy AI	Neuro-symbolic AI.
OAM	Orbital angular momentum
OFDM	Orthogonal frequency division multiplexing.
O&M	Orchestration and management.
PDF	Probability density function.
PHY	Physical layer.
PHYSec	PHY security.
PMF	Probability mass function.
QoS	Quality of service.
RANs	Radio access networks.
RAT	Radio access technology.

ACKNOWLEDGMENT

The author Tilahun M. Getu acknowledges Dr. Hamid Gharavi (Life Fellow, IEEE) of the National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, for leadership and funding support. The authors acknowledge the anonymous reviewers whose criticisms have guided the improvement of their submission.

DISCLAIMER

The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[2] M. Hirzallah, M. Krunz, B. Kecicioglu, and B. Hamzeh, "5G new radio unlicensed: Challenges and evaluation," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 3, pp. 689–701, Sep. 2021.

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[4] C. Chaccour, M. Naderi Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low latency communications for wireless VR?" 2020, *arXiv:2005.00536*.

[5] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[6] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 836–886, 2021.

[7] M. Alsabah, M. A. Naser, B. M. Mahmmod, S. H. Abdulhussain, M. R. Eissa, A. Al-Baidhani, N. K. Noordin, S. M. Sait, K. A. Al-Utaibi, and F. Hashim, "6G wireless communications networks: A comprehensive survey," *IEEE Access*, vol. 9, pp. 148191–148243, 2021.

[8] X. You, C. X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang, and M. Zhu, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, pp. 1–74, Jan. 2021.

[9] Y. Hao, Y. Miao, M. Chen, H. Gharavi, and V. Leung, "6G cognitive information theory: A mailbox perspective," *Big Data Cognit. Comput.*, vol. 5, no. 4, p. 56, Oct. 2021.

[10] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[11] P. Porambage, G. Gür, D. P. M. Osorio, M. Liyanage, A. Gurtov, and M. Ylianttila, "The roadmap to 6G security and privacy," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1094–1122, 2021.

[12] J. R. Bhat and S. A. Alqahtani, "6G ecosystem: Current status and future perspective," *IEEE Access*, vol. 9, pp. 43134–43167, 2021.

[13] A. Shahraki, M. Abbasi, M. Jalil Piran, and A. Taherkordi, "A comprehensive survey on 6G networks: Applications, core services, enabling technologies, and future challenges," 2021, *arXiv:2101.12475*.

[14] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133995–134030, 2020.

[15] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electron.*, vol. 3, no. 1, pp. 20–29, Jan. 2020.

[16] Y. Lu and X. Zheng, "6G: A survey on technologies, scenarios, challenges, and the related issues," *J. Ind. Inf. Integr.*, vol. 19, Sep. 2020, Art. no. 100158.

[17] E. Yaacoub and M.-S. Alouini, "A key 6G challenge and opportunity—Connecting the base of the pyramid: A survey on rural connectivity," *Proc. IEEE*, vol. 108, no. 4, pp. 533–582, Apr. 2020.

[18] J. Zhao, "A survey of intelligent reflecting surfaces (IRSs): Towards 6G wireless communication networks," 2019, *arXiv:1907.04789*.

[19] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 957–975, 2020.

[20] H. Viswanathan and P. E. Mogensen, "Communications in the 6G era," *IEEE Access*, vol. 8, pp. 57063–57074, 2020.

[21] L. Bariah, L. Mohjazi, S. Muhaidat, P. C. Sofotasios, G. K. Kurt, H. Yanikomeroglu, and O. A. Dobre, "A prospective look: Key enabling technologies, applications and open research topics in 6G networks," *IEEE Access*, vol. 8, pp. 174792–174820, 2020.

[22] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[23] M. A. Uusitalo, P. Rugeland, M. R. Boldi, E. C. Strinati, P. Demestichas, M. Ericson, G. P. Fettweis, M. C. Filippou, A. Gati, M.-H. Hamon, M. Hoffmann, M. Latva-Aho, A. Pärssinen, B. Richerzhagen, H. Schotten, T. Svensson, G. Wikström, H. Wymeersch, V. Ziegler, and Y. Zou, "6G vision, value, use cases and technologies from European 6G flagship project Hexa-X," *IEEE Access*, vol. 9, pp. 160004–160020, 2021.

[24] G. P. Fettweis and H. Boche, "6G: The personal tactile internet—And open questions for information theory," *IEEE BITS Inf. Theory Mag.*, vol. 1, no. 1, pp. 71–82, Sep. 2021.

[25] C. De Lima, D. Belot, R. Berkvens, A. Bourdoux, D. Dardari, M. Guillaud, M. Isomursu, E.-S. Lohan, Y. Miao, A. N. Barreto, M. R. K. Aziz, J. Saloranta, T. Sanguanpuak, H. Sarieddeen, G. Seco-Granados, J. Suutala, T. Svensson, M. Valkama, B. Van Liempd, and H. Wymeersch, "Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges," *IEEE Access*, vol. 9, pp. 26902–26925, 2021.

[26] L. U. Khan, I. Yaqoob, M. Imran, Z. Han, and C. S. Hong, "6G wireless systems: A vision, architectural elements, and future directions," *IEEE Access*, vol. 8, pp. 147029–147044, 2020.

[27] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[28] S. Chen, Y.-C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, "Vision, requirements, and technology trend of 6G: How to tackle the challenges of system coverage, capacity, user data-rate and movement speed," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 218–228, Apr. 2020.

[29] E. Bertin, N. Crespi, and T. Magedanz, *Shaping Future 6G Networks: Needs, Impacts, and Technologies*. Hoboken, NJ, USA: Wiley, 2022.

[30] P. Popovski, F. Chiariotti, V. Croisfelt, A. E. Kalør, I. Leyva-Mayorga, L. Marchegiani, S. Raj Pandey, and B. Soret, "Internet of Things (IoT) connectivity in 6G: An interplay of time, space, intelligence, and value," 2021, *arXiv:2111.05811*.

[31] M. Maier, A. Ebrahimzadeh, S. Rostami, and A. Beniiche, "The Internet of No Things: Making the internet disappear and 'see the invisible,'" *IEEE Commun. Mag.*, vol. 58, no. 11, pp. 76–82, Nov. 2020.

[32] M. Xu, W. C. Ng, W. Y. B. Lim, J. Kang, Z. Xiong, D. Niyato, Q. Yang, X. S. Shen, and C. Miao, "A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges," 2022, *arXiv:2203.05471*.

[33] N. Rajatheva, I. Atzeni, S. Bicais, E. Bjornson, A. Bourdoux, S. Buzzi, C. D'Andrea, J. B. Dore, S. Erkucuk, M. Fuentes, and K. Guan, "Scoring the terabit/s goal: Broadband connectivity in 6G," 2020, *arXiv:2008.07220*.

[34] N. Rajatheva, I. Atzeni, E. Bjornson, A. Bourdoux, S. Buzzi, J.-B. Dore, S. Erkucuk, M. Fuentes, K. Guan, Y. Hu, and X. Huang, "White paper on broadband connectivity in 6G," 2020, *arXiv:2004.14247*.

[35] F. Tariq, M. Khandaker, K.-K. Wong, M. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," 2019, *arXiv:1902.06700*.

[36] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: Opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 126–132, Oct. 2020.

[37] A. Celik, B. Shihada, and M.-S. Alouini, "Wireless data center networks: Advances, challenges, and opportunities," 2018, *arXiv:1811.11717*.

[38] B. Zong, C. Fan, X. Wang, X. Duan, B. Wang, and J. Wang, "6G technologies: Key drivers, core requirements, system architectures, and enabling technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 18–27, Sep. 2019.

[39] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Towards 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020.

[40] M. Katz, M. Matinmikko-Blue, and M. Latva-Aho, "6Genesis flagship program: Building the bridges towards 6G-enabled wireless smart society and ecosystem," in *Proc. IEEE 10th Latin-American Conf. Commun. (LATINCOM)*, Nov. 2018, pp. 1–9.

[41] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6G wireless communications: Vision and potential techniques," *IEEE Netw.*, vol. 33, no. 4, pp. 70–75, Jul. 2019.

[42] E. Calvanese Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, Sep. 2019.

[43] T. M. Getu, G. Kaddoum, and M. Bennis, "Making sense of meaning: A survey on metrics for semantic and goal-oriented communication," *IEEE Access*, vol. 11, pp. 45456–45492, 2023.

[44] L. Zhang, Y.-C. Liang, and D. Niyato, "6G visions: Mobile ultra-broadband, super Internet-of-Things, and artificial intelligence," *China Commun.*, vol. 16, no. 8, pp. 1–14, Aug. 2019.

[45] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[46] H. Gacanin and M. Di Renzo, "Wireless 2.0: Towards an intelligent radio environment empowered by reconfigurable meta-surfaces and artificial intelligence," 2020, *arXiv:2002.11040*.

[47] M. D. Renzo, M. Debbah, D.-T. Phan-Huy, A. Zappone, M.-S. Alouini, C. Yuen, V. Sciancalepore, G. C. Alexandropoulos, J. Hoydis, H. Gacanin, J. D. Rosny, A. Bounceur, G. Lerosey, and M. Fink, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, Dec. 2019.

[48] X. Huang, J. A. Zhang, R. P. Liu, Y. J. Guo, and L. Hanzo, "Airplane-aided integrated networking for 6G wireless: Will it work?" *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 84–91, Sep. 2019.

[49] S. J. Nawaz, S. K. Sharma, S. Wyne, M. N. Patwary, and M. Asaduzzaman, "Quantum machine learning for 6G communication networks: State-of-the-art and vision for the future," *IEEE Access*, vol. 7, pp. 46317–46350, 2019.

[50] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next? Five promising research directions for antenna arrays," 2019, *arXiv:1902.07678*.

[51] Y. Xiao, G. Shi, Y. Li, W. Saad, and H. V. Poor, "Toward self-learning edge intelligence in 6G," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 34–40, Dec. 2020.

[52] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[53] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proc. IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[54] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, "Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 60–69, Sep. 2019.

[55] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019.

[56] H. Gacanin, "Autonomous wireless systems with artificial intelligence: A knowledge management perspective," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 51–59, Sep. 2019.

[57] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[58] H.-C. Yang and M.-S. Alouini, "Data-oriented transmission in future wireless systems: Toward trustworthy support of advanced Internet of Things," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 78–83, Sep. 2019.

[59] E. Calvanese Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," 2020, *arXiv:2011.14844*.

[60] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 1st Quart., 2023.

[61] T. M. Getu, G. Kaddoum, and M. Bennis, "Tutorial-cum-survey on semantic and goal-oriented communication: Research landscape, challenges, and future directions," 2023, *arXiv:2308.01913*.

[62] E. Peltonen, M. Bennis, M. Capobianco, M. Debbah, A. Ding, F. Gil-Castiñeira, M. Jurmu, T. Karvonen, M. Kelanti, A. Kliks, T. Leppänen, L. Lovén, T. Mikkonen, A. Rao, S. Samarakoon, K. Seppänen, P. Sroka, S. Tarkoma, and T. Yang, "6G white paper on edge intelligence," 2020, *arXiv:2004.14850*.

[63] S. Zhang, J. Liu, H. Guo, M. Qi, and N. Kato, "Envisioning device-to-device communications in 6G," *IEEE Netw.*, vol. 34, no. 3, pp. 86–91, May 2020.

[64] B. A. Coll-Perales, J. Gozalvez, and J. L. Maestre, "5G and beyond: Smart devices as part of the network fabric," *IEEE Netw.*, vol. 33, no. 4, pp. 170–177, Jul. 2019.

[65] Y. Al-Eryani and E. Hossain, "The D-OMA method for massive multiple access in 6G: Performance, security, and challenges," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 92–99, Sep. 2019.

[66] E. Basar, "Reconfigurable intelligent surface-based index modulation: A new beyond MIMO paradigm for 6G," 2019, *arXiv:1904.06704*.

[67] T. M. Getu, W. Saad, G. Kaddoum, and M. Bennis, "Performance limits of a deep learning-enabled text semantic communication under interference," 2023, *arXiv:2302.14702*.

[68] T. M. Getu, G. Kaddoum, and M. Bennis, "Semantic communication: A survey on research landscape, challenges, and future directions," *TechRxiv Preprint*, 2023, doi: 10.36227/techrxiv.24527455.v1.

[69] T. M. Getu, G. Kaddoum, and M. Bennis, "Deep learning-enabled text semantic communication under interference: An empirical study," 2023, *arXiv:2310.19974*.

[70] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL, USA: Univ. Illinois Press, 1949.

[71] Y. Liu, X. Wang, Z. Ning, M. Zhou, L. Guo, and B. Jedari, "A survey on semantic communications: Technologies, solutions, applications and challenges," *Digit. Commun. Netw.*, Jun. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864823000925?via%3Dihub

[72] W. Yang, H. Du, Z. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. S. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," 2022, *arXiv:2207.00427*.

[73] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 210–219, Feb. 2022.

[74] P. Zhang, W. Xu, H. Gao, K. Niu, X. Xu, X. Qin, C. Yuan, Z. Qin, H. Zhao, J. Wei, and F. Zhang, "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks," *Engineering*, vol. 8, pp. 60–73, Jan. 2022.

[75] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Ye Li, "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.

[76] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, Dec. 2021.

[77] Z. Lu, R. Li, K. Lu, X. Chen, E. Hossain, Z. Zhao, and H. Zhang, "Semantics-empowered communication: A tutorial-cum-survey," 2022, *arXiv:2212.08487*.

[78] D. Gunduz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," 2022, *arXiv:2207.09353*.

[79] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," 2022, *arXiv:2211.14343*.

[80] D. Wheeler and B. Natarajan, "Engineering semantic communication: A survey," 2022, *arXiv:2208.06314*.

[81] S. Iyer, R. Khanai, D. Torse, R. J. Pandya, K. Rabie, K. Pai, W. U. Khan, and Z. Fadlullah, "A survey on semantic communications for intelligent wireless networks," 2022, *arXiv:2202.03705*.

[82] F. Zhou, Y. Li, X. Zhang, Q. Wu, X. Lei, and R. Q. Hu, "Cognitive semantic communication systems driven by knowledge graph," 2022, *arXiv:2202.11958*.

[83] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 96–102, Jun. 2021.

[84] H. Seo, J. Park, M. Bennis, and M. Debbah, "Semantics-native communication with contextual reasoning," 2021, *arXiv:2108.05681*.

[85] C. Zhang, H. Zou, S. Lasaulce, W. Saad, M. Kountouris, and M. Bennis, "Goal-oriented communications for the IoT and application to data compression," 2022, *arXiv:2211.05378*.

[86] M. Goek, "Semantic and goal-oriented signal processing: Semantic extraction," M.S. thesis, Graduate School Eng. Sci., Bilkent Univ., Turkey, Aug. 2022.

[87] S. Xie, S. Ma, M. Ding, Y. Shi, M. Tang, and Y. Wu, "Robust information bottleneck for task-oriented communication with digital modulation," 2022, *arXiv:2209.10382*.

[88] A. Mostaani, T. X. Vu, S. K. Sharma, V.-D. Nguyen, Q. Liao, and S. Chatzinotas, "Task-oriented communication design in cyber-physical systems: A survey on theory and applications," 2021, *arXiv:2102.07166*.

[89] P. H. Pathak, X. Feng, P. Hu, and P. Mohapatra, "Visible light communication, networking, and sensing: A survey, potential and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2047–2077, 4th Quart., 2015.

[90] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[91] N. Kato, B. Mao, F. Tang, Y. Kawamoto, and J. Liu, "Ten challenges in advancing machine learning technologies toward 6G," *IEEE Wireless Commun.*, vol. 27, no. 3, pp. 96–103, Jun. 2020.

[92] M. Ylianttila, R. Kantola, A. Gurtov, L. Mucchi, I. Oppermann, Z. Yan, T. H. Nguyen, F. Liu, T. Hewa, M. Liyanage, and A. Ijaz, "6G white paper: Research challenges for trust, security and privacy," 2020, *arXiv:2004.11665*.

[93] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6G: Machine-learning approaches," *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.

[94] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart., 2018.

[95] N. Kato, Z. M. Fadlullah, F. Tang, B. Mao, S. Tani, A. Okamura, and J. Liu, "Optimizing space-air-ground integrated networks by artificial intelligence," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 140–147, Aug. 2019.

[96] I. Boybat, M. Le Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, "Neuromorphic computing with multi-memristive synapses," *Nature Commun.*, vol. 9, no. 1, p. 2514, Jun. 2018.

[97] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas, *Event Based Neuromorphic Systems*. Hoboken, NJ, USA: Wiley, 2015.

[98] B. Rajendran, A. Sebastian, M. Schmuker, N. Srinivasa, and E. Eleftheriou, "Low-power neuromorphic hardware for signal processing applications," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 1–97, Nov. 2019.

[99] N. Soures and D. Kudithipudi, "Spiking reservoir networks: Brain-inspired recurrent algorithms that use random, fixed synaptic strengths," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 78–87, Nov. 2019.

[100] C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, "A survey of neuromorphic computing and neural networks in hardware," 2017, *arXiv:1705.06963*.

[101] K. David and H. Berndt, "6G vision and requirements: Is there any need for beyond 5G?" *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 72–80, Sep. 2018.

[102] M. Latva-Aho and K. Leppanen. (Sep. 2019). *Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence*. [Online]. Available: http://jultika.oulu.fi/Record/isbn978-952-62-2354-4

[103] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[104] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, 2014.

[105] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[106] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M. L. Shyu, S. C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, pp. 1–36, Sep. 2018.

[107] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," 2018, *arXiv:1809.02165*.

[108] T. G. Dietterich, "Steps toward robust artificial intelligence," *AI Mag.*, vol. 38, no. 3, pp. 3–24, Sep. 2017.

[109] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY, USA: Pantheon, 2019.

[110] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2018.

[111] M. I. Jordan. (Jul. 2019). *Artificial Intelligence—The Revolution Hasn't Happened Yet*. Harvard Data Science Review. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9

[112] Stanford University. (2016). *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence (AI100)*. [Online]. Available: https://ai100.stanford.edu

[113] S. S. Adams, I. Arel, J. Bach, R. Coop, R. Furlan, B. Goertzel, J. S. Hall, A. Samsonovich, M. Scheutz, M. Schlesinger, S. C. Shapiro, and J. F. Sowa, "Mapping the landscape of human-level artificial general intelligence," *AI Mag.*, vol. 33, pp. 25–42, Jan. 2012.

[114] M. T. Bennett and Y. Maruyama, "The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence," 2021, *arXiv:2110.01831*.

[115] A. Franz, "Artificial general intelligence through recursive data compression and grounded reasoning: A position paper," 2015, *arXiv:1506.04366*.

[116] B. Goertzel, "Artificial general intelligence: Concept, state of the art, and future prospects," *J. Artif. Gen. Intell.*, vol. 5, no. 1, pp. 1–48, Dec. 2014.

[117] A. M. Khalili. (2022). *Artificial General Intelligence: A New Perspective, With Application to Scientific Discovery*. [Online]. Available: https://engrxiv.org/preprint/view/732/4372

[118] H. Yamakawa, M. Osawa, and Y. Matsuo, "Whole brain architecture approach is a feasible way toward an artificial general intelligence," in *Proc. ICONIP*, 2016, pp. 275–281.

[119] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, H. Sun, and J.-R. Wen, "Towards artificial general intelligence via a multimodal foundation model," *Nature Commun.*, vol. 13, no. 1, p. 3094, Jun. 2022.

[120] H. Yamakawa, "The whole brain architecture approach: Accelerating the development of artificial general intelligence by referring to the brain," *Neural Netw.*, vol. 144, pp. 478–495, Dec. 2021.

[121] J. Clune, "AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence," 2019, *arXiv:1905.10985*.

[122] B. Goertzel, "The general theory of general intelligence: A pragmatic patternist perspective," 2021, *arXiv:2103.15100*.

[123] R. Fjelland, "Why general artificial intelligence will not be realized," *Humanities Social Sci. Commun.*, vol. 7, no. 1, pp. 1–9, Jun. 2020.

[124] Z. Chen and B. Liu, "Lifelong machine learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 10, no. 3, pp. 1–145, Nov. 2016.

[125] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," 2018, *arXiv:1802.07569*.

[126] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.

[127] G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature Commun.*, vol. 11, no. 1, p. 4069, Aug. 2020.

[128] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," 2019, *arXiv:1907.00182*.

[129] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," 2020, *arXiv:2010.15277*.

[130] D. L. Silver, Q. Yang, and L. Li, "Lifelong machine learning systems: Beyond learning algorithms," in *Proc. AAAI*, 2013, pp. 1–7.

[131] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, and J. Krishnamurthy, "Never-ending learning," *Commun. ACM*, vol. 61, no. 5, pp. 103–115, 2018.

[132] C. Finn, "Learning to learn with gradients," Ph.D. dissertation, EECS Dept., Univ. California, Berkeley, CA, USA, Aug. 2018. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-105.html

[133] J. Vanschoren, "Meta-learning: A survey," 2018, *arXiv:1810.03548*.

[134] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[135] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, May 2015.

[136] J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[137] J. Pearl, *Causality: Models, Reasoning Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[138] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 27, pp. 7345–7352, Jul. 2016.

[139] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.

[140] B. Schölkopf, "Causality for machine learning," 2019, *arXiv:1911.10500*.

[141] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal machine learning: A survey and open problems," 2022, *arXiv:2206.15475*.

[142] E. Davis, "Logical formalizations of commonsense reasoning: A survey," *J. Artif. Intell. Res.*, vol. 59, pp. 651–723, Aug. 2017.

[143] E. Davis and G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence," *Commun. ACM*, vol. 58, no. 9, pp. 92–103, Aug. 2015.

[144] T. R. Besold, A. D. Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kuehnberger, L. C. Lamb, D. Lowd, P. M. V. Lima, L. de Penning, G. Pinkas, H. Poon, and G. Zaverucha, "Neural-symbolic learning and reasoning: A survey and interpretation," 2017, *arXiv:1711.03902*.

[145] A. Garcez, L. C. Lamb, and D. M. Gabbay, *Neural-Symbolic Cognitive Reasoning*. Berlin, Germany: Springer, 2009.

[146] L. de Penning, A. Garcez, L. C. Lamb, and J. J. Meyer, "A neural-symbolic cognitive agent for online learning and reasoning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1653–1658.

[147] A. D. Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd wave," 2020, *arXiv:2012.05876*.

[148] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[149] D. Paulius and Y. Sun, "A survey of knowledge representation in service robotics," *Robot. Auto. Syst.*, vol. 118, pp. 13–30, Aug. 2019.

[150] Z. Bouraoui, A. Cornuéjols, T. Denœux, S. Destercke, D. Dubois, R. Guillaume, J. Marques-Silva, J. Mengin, H. Prade, S. Schockaert, M. Serrurier, and C. Vrain, "From shallow to deep interactions between knowledge representation, reasoning and machine learning," 2019, *arXiv:1912.06612*.

[151] F. van Harmelen, V. Lifschitz, and B. Porter, *Handbook of Knowledge Representation* (Foundations of Artificial Intelligence), vol. 3. Amsterdam, The Netherlands: Elsevier, 2008.

[152] K. Trentelman. (Jul. 2009). *Survey of Knowledge Representation and Reasoning Systems*. [Online]. Available: https://apps.dtic.mil/sti/pdfs/ADA508761.pdf

[153] D. Harel, A. Marron, and J. Sifakis, "Autonomics: In search of a foundation for next-generation autonomous systems," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 30, pp. 17491–17498, Jul. 2020.

[154] J. Sifakis, "Autonomous systems—An architectural characterization," 2018, *arXiv:1811.10277*.

[155] D. Harel, A. Marron, and J. Sifakis, "Creating a foundation for next-generation autonomous systems," *IEEE Des. Test*, vol. 39, no. 1, pp. 49–56, Feb. 2022.

[156] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[157] G. Shi, D. Gao, X. Song, J. Chai, M. Yang, X. Xie, L. Li, and X. Li, "A new communication paradigm: From bit accuracy to semantic fidelity," 2021, *arXiv:2101.12649*.

[158] B. A. Juba and M. Sudan, "Universal semantic communication II: A theory of goal-oriented communication," *Electron. Colloq. Comput. Complex.*, vol. 15, Jan. 2008.

[159] M. Kalfa, M. Gok, A. Atalik, B. Tegin, T. M. Duman, and O. Arikan, "Towards goal-oriented semantic signal processing: Applications and future challenges," *Digit. Signal Process.*, vol. 119, Dec. 2021, Art. no. 103134.

[160] W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Cao, and K. B. Letaief, "Semantic communication meets edge intelligence," 2022, *arXiv:2202.06471*.

[161] M. Kalfa, S. Y. Yetim, A. Atalik, M. Gök, Y. Ge, R. Li, W. Tong, T. M. Duman, and O. Arikan, "Reliable extraction of semantic information and rate of innovation estimation for graph signals," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 119–140, Jan. 2023.

[162] 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.

[163] B. R. Barricelli, E. Casiraghi, and D. Fogli, "A survey on digital twin: Definitions, characteristics, applications, and design implications," *IEEE Access*, vol. 7, pp. 167653–167671, 2019.

[164] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019.

[165] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," *IEEE Access*, vol. 8, pp. 108952–108971, 2020.

[166] S. Boschert, T. Coughlin, M. Ferraris, F. Flammini, J. G. Florido, A. C. Gonzalez, P. Henz, D. Kerckhove, R. Rosen, R. Saracco, and A. Singh. (2019). *Symbiotic Autonomous Systems*. Acessed: Jan. 2021. [Online]. Available: https://digitalreality.ieee.org/images/files/pdf/1SAS_WP3_Nov2019.pdf

[167] *Network 2030—Additional Representative Use Cases and Key Network Requirements for Network 2030*, Standard ITU-T-FG-NET2030, Jun. 2020.

[168] S. Nayak and R. Patgiri, "6G communication technology: A vision on intelligent healthcare," 2020, *arXiv:2005.07532*.

[169] P. Di Lorenzo, M. Merluzzi, F. Binucci, C. Battiloro, P. Banelli, E. C. Strinati, and S. Barbarossa, "Goal-oriented communications for the IoT: System design and adaptive resource optimization," 2023, *arXiv:2310.13948*.

[170] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 319–352, 1st Quart., 2023.

[171] W. Yang Bryan Lim, Z. Xiong, D. Niyato, X. Cao, C. Miao, S. Sun, and Q. Yang, "Realizing the metaverse with edge intelligence: A match made in heaven," 2022, *arXiv:2201.01634*.

[172] H. Zhou, X. Liu, Y. Deng, N. Pappas, and A. Nallanathan, "Task-oriented and semantics-aware 6G networks," 2022, *arXiv:2210.09372*.

[173] D. Wen, P. Liu, G. Zhu, Y. Shi, J. Xu, Y. C. Eldar, and S. Cui, "Task-oriented sensing, computation, and communication integration for multi-device edge AI," *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2486–2502, Mar. 2024.

[174] S. Ma, W. Qiao, Y. Wu, H. Li, G. Shi, D. Gao, Y. Shi, S. Li, and N. Al-Dhahir, "Task-oriented explainable semantic communications," 2023, *arXiv:2302.13560*.

[175] W. Xu, Y. Zhang, F. Wang, Z. Qin, C. Liu, and P. Zhang, "Semantic communication for Internet of Vehicles: A multi-user cooperative approach," 2022, *arXiv:2212.03037*.

[176] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sep. 2022.

[177] C. Kurisummoottil Thomas and W. Saad, "Neuro-symbolic artificial intelligence (AI) for intent based semantic communication," 2022, *arXiv:2205.10768*.

[178] Y. Yang, C. Guo, F. Liu, C. Liu, L. Sun, Q. Sun, and J. Chen, "Semantic communications with AI tasks," 2021, *arXiv:2109.14170*.

[179] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869–904, 2nd Quart., 2020.

[180] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[181] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.

[182] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[183] J. Wu, R. Li, X. An, C. Peng, Z. Liu, J. Crowcroft, and H. Zhang, "Toward native artificial intelligence in 6G networks: System design, architectures, and paradigms," 2021, *arXiv:2103.02823*.

[184] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: The confluence of edge computing and artificial intelligence," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7457–7469, Aug. 2020.

[185] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 45–66, 2020.

[186] P. Carbone, G. Dan, J. Gross, B. Goeransson, and M. Petrova, "NeuroRAN: Rethinking virtualization for AI-native radio access networks in 6G," 2021, *arXiv:2104.08111*.

[187] J. Hoydis, F. A. Aoudia, A. Valcarce, and H. Viswanathan, "Toward a 6G AI-native air interface," *IEEE Commun. Mag.*, vol. 59, no. 5, pp. 76–81, May 2021.

[188] W. Tong and P. Zhu, *6G: The Next Horizon: From Connected People and Things to Connected Intelligence*. Cambridge, U.K.: Cambridge Univ. Press, 2021.

[189] I. F. Akyildiz and A. Kak, "The Internet of Space Things/CubeSats: A ubiquitous cyber-physical system for the connected world," *Comput. Netw.*, vol. 150, pp. 134–149, Feb. 2019.

[190] K. M. S. Huq, S. A. Busari, J. Rodriguez, V. Frascolla, W. Bazzi, and D. C. Sicker, "Terahertz-enabled wireless system for beyond-5G ultra-fast networks: A brief survey," *IEEE Netw.*, vol. 33, no. 4, pp. 89–95, Jul. 2019.

[191] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, P. Popovski, and M. Debbah, "Seven defining features of terahertz (THz) wireless systems: A fellowship of communication and sensing," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 967–993, 2nd Quart., 2022.

[192] K Tekbiyik, A. R. Ekti, G. K. Kurt, and A. Görçin, "Terahertz band communication systems: Challenges, novelties and standardization efforts," *Phys. Commun.*, vol. 35, Aug. 2019, Art. no. 100700.

[193] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "User-centric cell-free massive MIMO networks: A survey of opportunities, challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 611–652, 1st Quart., 2022.

[194] P. Popovski, O. Simeone, F. Boccardi, D. Gunduz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5G wireless connectivity," 2019, *arXiv:1907.02441*.

[195] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, T. Soleymani, B. Soret, and K. Henrik Johansson, "Semantic communications in networked systems: A data significance perspective," 2021, *arXiv:2103.05391*.

[196] M. Sana and E. C. Strinati, "Learning semantics: An opportunity for effective 6G communications," 2021, *arXiv:2110.08049*.

[197] E. G. Soyak and O. Ercetin, "Effective communications for 6G: Challenges and opportunities," 2022, *arXiv:2203.11695*.

[198] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6G: Vision, principles, and technologies," 2023, *arXiv:2303.10920*.

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

[200] J. Dommel, D. Wieruch, Z. Utkovski, and S. Stanczak, "A semantics-aware communication scheme to estimate the empirical measure of a quantity of interest via multiple access fading channels," in *Proc. IEEE Stat. Signal Process. Workshop*, Jun. 2021, pp. 521–525.

[201] J. Guo and C. Yang, "Learning precoding for semantic communications," in *Proc. ICC Workshops*, 2022, pp. 163–168.

[202] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5181–5192, Aug. 2022.

[203] M. Karimzadeh Farshbafan, W. Saad, and M. Debbah, "Common language for goal-oriented semantic communications: A curriculum learning framework," 2021, *arXiv:2111.08051*.

[204] M. Karimzadeh Farshbafan, W. Saad, and M. Debbah, "Curriculum learning for goal-oriented semantic communications with a common language," 2022, *arXiv:2204.10429*.

[205] C. Wang, X. Yu, L. Xu, Z. Wang, and W. Wang, "Multimodal semantic communication accelerated bidirectional caching for 6G MEC," *Future Gener. Comput. Syst.*, vol. 140, pp. 225–237, Mar. 2023.

[206] J. Kang, H. Du, Z. Li, Z. Xiong, S. Ma, D. Niyato, and Y. Li, "Personalized saliency in task-oriented semantic communications: Image transmission and performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 186–201, Jan. 2023.

[207] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[208] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv:0004057*.

[209] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 73–87, Jan. 2023.

[210] I. E. Aguerri and A. Zaidi, "Distributed variational representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 120–138, Jan. 2021.

[211] J. Shao, X. Zhang, and J. Zhang, "Task-oriented communication for edge video analytics," 2022, *arXiv:2211.14049*.

[212] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," 2016, *arXiv:1604.00268*.

[213] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gündüz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2590–2603, Aug. 2021.

[214] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *Proc. IEEE Int. Conf. Auto. Syst. (ICAS)*, Aug. 2021, pp. 1–5.

[215] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," 2022, *arXiv:2209.07689*.

[216] G. Zhang, Q. Hu, Z. Qin, Y. Cai, and G. Yu, "A unified multi-task semantic communication system with domain adaptation," 2022, *arXiv:2206.00254*.

[217] H. Xie, Z. Qin, and G. Ye Li, "Task-oriented multi-user semantic communications for VQA task," 2021, *arXiv:2108.07357*.

[218] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, "Dynamic resource allocation for multi-user goal-oriented communications at the wireless edge," in *Proc. EUSIPCO*, 2022, pp. 697–701.

[219] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, "Adaptive resource optimization for edge inference with goal-oriented communications," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, p. 123, Dec. 2022.

[220] C. Liu, C. Guo, Y. Yang, and N. Jiang, "Adaptable semantic compression and resource allocation for task-oriented communications," 2022, *arXiv:2204.08910*.

[221] C. Liu, C. Guo, Y. Yang, and J. Chen, "Bandwidth and power allocation for task-oriented SemanticCommunication," 2022, *arXiv:2201.10795*.

[222] M. Merluzzi, M. C. Filippou, L. G. Baltar, M. D. Muek, and E. C. Strinati, "6G goal-oriented communications: How to coexist with legacy systems?" 2023, *arXiv:2308.13358*.

[223] R. Ahlswede and G. Dueck, "Identification via channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 15–29, Jan. 1989.

[224] R. Ahlswede and G. Dueck, "Identification in the presence of feedback—A discovery of new capacity formulas," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 30–36, Jan. 1989.

[225] R. Ahlswede and Z. Zhang, "New directions in the theory of identification via channels," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1040–1050, Jul. 1995.

[226] E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio, "Flow network based generative models for non-iterative diverse candidate generation," in *Proc. NIPS*, 2021, pp. 1–14.

[227] S. Derebeyoglu, C. Deppe, and R. Ferrara, "Performance analysis of identification codes," *Entropy*, vol. 22, no. 10, p. 1067, Sep. 2020.

[228] C. V. Lengerke, A. Hefele, J. A. Cabrera, O. Kosut, M. Reisslein, and F. H. P. Fitzek, "Identification codes: A topical review with design guidelines for practical systems," *IEEE Access*, vol. 11, pp. 14961–14982, 2023.

[229] C. von Lengerke, A. Hefele, J. A. Cabrera, M. Reisslein, and F. H. P. Fitzek, "Beyond the bound: A new performance perspective for identification via channels," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2687–2706, Aug. 2023.

[230] A. Krechowicz, S. Deniziak, and G. Lukawski, "Highly scalable distributed architecture for NoSQL datastore supporting strong consistency," *IEEE Access*, vol. 9, pp. 69027–69043, 2021.

[231] M. Pedrosa, R. Lebre, and C. Costa, "A performant protocol for distributed health records databases," *IEEE Access*, vol. 9, pp. 125930–125940, 2021.

[232] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[233] Y. Xu, Z. Hui, and Y. Deng, "Task-oriented semantics-aware communication for wireless UAV control and command transmission," 2023, *arXiv:2306.14228*.

[234] T. Rausch and S. Dustdar, "Edge intelligence: The convergence of humans, things, and AI," in *Proc. IEEE Int. Conf. Cloud Eng. (IC2E)*, Jun. 2019, pp. 86–96.

[235] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[236] A. D. Raha, A. Adhikary, S. Dam, S.-B. Park, and C. Hong. (Dec. 2022). *A Goal-Oriented Semantic Communication Framework for Connected and Autonomous Vehicular Network: A Deep Auto-Encoder Approach*. [Online]. Available: https://www.researchgate.net/publication/366514075_A_Goal-Oriented_Semantic_Communication_Framework_for_Connected_and_Autonomous_Vehicular_Network_A_Deep_Auto-Encoder_Approach

[237] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 230–244, Jan. 2023.

[238] B. Kizilkaya, C. She, G. Zhao, and M. A. Imran, "Task-oriented prediction and communication co-design for haptic communications," *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 8987–9001, Jul. 2023.

[239] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Toward haptic communications over the 5G tactile internet," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3034–3059, 4th Quart., 2018.

[240] D. Van Den Berg, R. Glans, D. De Koning, F. A. Kuipers, J. Lugtenburg, K. Polachan, P. T. Venkata, C. Singh, B. Turkovic, and B. Van Wijk, "Challenges in haptic communications over the tactile internet," *IEEE Access*, vol. 5, pp. 23502–23518, 2017.

[241] E. Steinbach, S. Hirche, M. Ernst, F. Brandi, R. Chaudhari, J. Kammerl, and I. Vittorias, "Haptic communications," *Proc. IEEE*, vol. 100, no. 4, pp. 937–956, Apr. 2012.

[242] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[243] N. Promwongsa, A. Ebrahimzadeh, D. Naboulsi, S. Kianpisheh, F. Belqasmi, R. Glitho, N. Crespi, and O. Alfandi, "A comprehensive survey of the tactile internet: State-of-the-art and research directions," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 472–523, 1st Quart., 2021.

[244] M. Maier, M. Chowdhury, B. P. Rimal, and D. P. Van, "The tactile internet: Vision, recent progress, and open challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 138–145, May 2016.

[245] 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*.

[246] S. K. Jagatheesaperumal, Z. Yang, Q. Yang, C. Huang, W. Xu, M. Shikh-Bahaei, and Z. Zhang, "Semantic-aware digital twin for metaverse: A comprehensive review," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 38–46, Aug. 2023.

[247] C. Hackl, D. Lueth, and T. Di Bartolo, *Navigating the Metaverse: A Guide to Limitless Possibilities in a Web 3.0 World*. Hoboken, NJ, USA: Wiley, 2022.

[248] H. Wang, H. Ning, Y. Lin, W. Wang, S. Dhelim, F. Farha, J. Ding, and M. Daneshmand, "A survey on Metaverse: The state-of-the-art, technologies, applications, and challenges," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14671–14688, Aug. 2023.

[249] S.-M. Park and Y.-G. Kim, "A metaverse: Taxonomy, components, applications, and open challenges," *IEEE Access*, vol. 10, pp. 4209–4251, 2022.

[250] L.-H. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, and P. Hui, "All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda," 2021, *arXiv:2110.05352*.

[251] T. Huynh-The, Q.-V. Pham, X.-Q. Pham, T. T. Nguyen, Z. Han, and D.-S. Kim, "Artificial intelligence for the metaverse: A survey," 2022, *arXiv:2202.10336*.

[252] J. Sun, W. Gan, H.-C. Chao, and P. S. Yu, "Metaverse: Survey, applications, security, and opportunities," 2022, *arXiv:2210.07990*.

[253] M. Ali, F. Naeem, G. Kaddoum, and E. Hossain, "Metaverse communications, networking, security, and applications: Research issues, state-of-the-art, and future directions," 2022, *arXiv:2212.13993*.

[254] A. Abilkaiyrkyzy, A. Elhagry, F. Laamarti, and A. E. Saddik, "Metaverse key requirements and platforms survey," *IEEE Access*, vol. 11, pp. 117765–117787, 2023.

[255] O. Hashash, C. Chaccour, W. Saad, T. Yu, K. Sakaguchi, and M. Debbah, "The seven worlds and experiences of the wireless metaverse: Challenges and opportunities," 2023, *arXiv:2304.10282*.

[256] Z. Meng, C. She, G. Zhao, M. A. Imran, M. Dohler, Y. Li, and B. Vucetic, "Task-oriented metaverse design in the 6G era," 2023, *arXiv:2306.03158*.

[257] K. Li, B. P. L. Lau, X. Yuan, W. Ni, M. Guizani, and C. Yuen, "Toward ubiquitous semantic Metaverse: Challenges, approaches, and opportunities," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21855–21872, Dec. 2023.

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

[259] K. Li, B. P. L. Lau, X. Yuan, W. Ni, M. Guizani, and C. Yuen, "Toward ubiquitous semantic metaverse: Challenges, approaches, and opportunities," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21855–21872, Dec. 2023.

[260] J. Chen, J. Wang, C. Jiang, Y. Ren, and L. Hanzo, "Trust-worthy semantic communications for the metaverse relying on federated learning," 2023, *arXiv:2305.09255*.

[261] M. Chehimi, O. Hashash, and W. Saad, "The roadmap to a quantum-enabled wireless metaverse: Beyond the classical limits," in *Proc. 5th Int. Conf. Adv. Comput. Tools Eng. Appl. (ACTEA)*, Jul. 2023, pp. 7–12.

[262] Y. Lin, H. Du, D. Niyato, J. Nie, J. Zhang, Y. Cheng, and Z. Yang, "Blockchain-aided secure semantic communication for AI-generated content in metaverse," *IEEE Open J. Comput. Soc.*, vol. 4, pp. 72–83, 2023.

[263] Y. Lin, Z. Gao, H. Du, D. Niyato, J. Kang, A. Jamalipour, and X. Sherman Shen, "A unified framework for integrating semantic communication and AI-generated content in metaverse," 2023, *arXiv:2305.11911*.

[264] Z. Wang, Y. Deng, and A. H. Aghvami, "Task-oriented and semantics-aware communication framework for augmented reality," 2023, *arXiv:2306.15470*.

[265] S. Rezwan, H. Wu, J. A. Cabrera, G. T. Nguyen, M. Reisslein, and F. H. P. Fitzek, "CXR+ voxel-based semantic compression for networked immersion," *IEEE Access*, vol. 11, pp. 52763–52777, 2023.

[266] Y. Zhu, Y. Huang, X. Qiao, Z. Tan, B. Bai, H. Ma, and S. Dustdar, "A semantic-aware transmission with adaptive control scheme for volumetric video service," *IEEE Trans. Multimedia*, vol. 25, pp. 7160–7172, 2023.

[267] Z. Yang, M. Chen, G. Li, Y. Yang, and Z. Zhang, "Secure semantic communications: Fundamentals and challenges," 2023, *arXiv:2301.01421*.

[268] G. Stamatakis, N. Pappas, A. Fragkiadakis, and A. Traganitis, "Semantics-aware active fault detection in status updating systems," 2022, *arXiv:2202.00923*.

[269] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[270] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 2nd ed. Belmont, MA, USA: Athena Scientific, 2008.

[271] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," 2015, *arXiv:1503.02406*.

[272] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[273] E. Beck, C. Bockelmann, and A. Dekorsy, "Semantic information recovery in wireless networks," 2022, *arXiv:2204.13366*.

[274] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20437–20448.

[275] P. A. Stavrou and M. Kountouris, "The role of fidelity in goal-oriented semantic communication: A rate distortion approach," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 3918–3931, Jul. 2023, doi: 10.1109/TCOMM.2023.3274122.

[276] P. A. Stavrou and M. Kountouris, "A rate distortion approach to goal-oriented communication," *TechRxiv Preprint*, 2022, doi: 10.36227/techrxiv.19128026.v1.

[277] F. Liu, W. Tong, Y. Yang, Z. Sun, and C. Guo, "Task-oriented image semantic communication based on rate-distortion theory," 2022, *arXiv:2201.10929*.

[278] H. Zou, C. Zhang, S. Lasaulce, L. Saludjian, and H. V. Poor, "Goal-oriented quantization: Analysis, design, and application to resource allocation," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 42–54, Jan. 2023.

[279] H. Zou, Y. Sun, C. Zhang, S. Lasaulce, M. Kieffer, and L. Saludjian, "Goal-oriented quantization: Applications to convex cost functions with polyhedral decision space," in *Proc. 20th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, Sep. 2022, pp. 291–297.

[280] J. Liu, S. Shao, W. Zhang, and H. V. Poor, "An indirect rate-distortion characterization for semantic sources: General model and the case of Gaussian observation," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 5946–5959, Sep. 2022.

[281] O. Goldreich, B. Juba, and M. Sudan, "A theory of goal-oriented communication," *J. ACM*, vol. 59, no. 2, pp. 1–65, Apr. 2012.

[282] Y. Zhong, "A theory of semantic information," *China Commun.*, vol. 14, no. 1, pp. 1–17, Jan. 2017.

[283] Y. Zhong and G. Dodig-Crnkovic,"A theory of semantic information in the context of its ecology," in *Theoretical Information Studies: Information in the World*, M. Burgin and G. Dodig-Crnkovic, Eds. Singapore: World Scientific, 2020, pp. 81–112

[284] W. Johannsen, "On semantic information in nature," *Information*, vol. 6, pp. 411–431, Jul. 2015.

[285] R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information," Res. Lab. Electron., MIT, Cambridge, MA, USA, 1952.

[286] J. Choi, S. W. Loke, and J. Park, "A unified view on semantic information and communication: A probabilistic logic approach," in *Proc. ICC Workshops*, 2022, pp. 705–710.

[287] N. J. Venhuizen, M. W. Crocker, and H. Brouwer, "Semantic entropy in language comprehension," *Entropy*, vol. 21, no. 12, p. 1159, Nov. 2019.

[288] I. D. Melamed, "Measuring semantic entropy," in *Proc. Workshop Tagging Text Lexical Semantics*, 1997. [Online]. Available: https://aclanthology.org/W97-0207/

[289] X. Liu, W. Jia, W. Liu, and W. Pedrycz, "AFSSE: An interpretable classifier with axiomatic fuzzy set and semantic entropy," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, pp. 2825–2840, Nov. 2020.

[290] W. Tong and G. Ye Li, "Nine challenges in artificial intelligence and wireless communications for 6G," 2021, *arXiv:2109.11320*.

[291] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[292] W. Guo, "Explainable artificial intelligence for 6G: Improving trust between human and machine," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 39–45, Jun. 2020.

[293] T. M. Getu, "Error bounds for a matrix-vector product approximation with deep ReLU neural networks," 2021, *arXiv:2111.12963*.

[294] T. Poggio, A. Banburski, and Q. Liao, "Theoretical issues in deep networks," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30039–30045, Jun. 2020.

[295] T. M. Getu and G. Kaddoum, "Fundamental limits of deep learning-based binary classifiers trained with Hinge loss," 2023, *arXiv:2309.06774*.

[296] T. M. Getu, N. T. Golmie, and D. W. Griffith, "Blind estimation of a doubly selective OFDM channel: A deep learning algorithm and theory," 2022, *arXiv:2206.07483*.

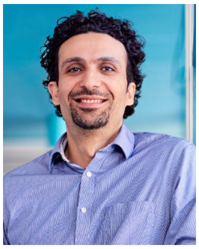
[297] A. Li, S. Wu, S. Meng, and Q. Zhang, "Towards goal-oriented semantic communications: New metrics, open challenges, and future research directions," 2023, *arXiv:2304.00848*.

[298] K. B. Carter, *Actionable Intelligence: A Guide to Delivering Business Results With Big Data Fast!* Hoboken, NJ, USA: Wiley, 2014.

TILAHUN M. GETU (Member, IEEE) received the Ph.D. degree in electrical engineering from École de Technologie Supérieure (ÉTS), Montreal, QC, Canada, in 2019.

He is currently a Postdoctoral Fellow with ÉTS. His research interests include the numerous fields of classical and quantum science, technology, engineering, and mathematics (STEM) at the nexus of communications, signal processing, and networking (all types); intelligence (both artificial and natural); robotics; computing; security; optimization; high-dimensional statistics; and high-dimensional causal inference. He has received several awards, including the 2019 ÉTS Board of Directors Doctoral Excellence Award in recognition of the Ph.D. dissertation selected as the 2019 ÉTS all-university best Ph.D. dissertation.

GEORGES KADDOUM (Senior Member, IEEE) is currently a Professor and a Tier 2 Canada Research Chair with École de Technologie Supérieure (ÉTS), Universite du Quebec, Montreal, Canada. He is also a Faculty Fellow with the Cyber Security Systems and Applied AI Research Center, Lebanese American University. His research interests include 5G/6G networks, tactical communications, resource allocations, and security. He has received many prestigious national and international awards in recognition of his outstanding research outcomes. He serves as an Area Editor for IEEE Transactions on Machine Learning in Communications and Networking and an Associate Editor for IEEE Transactions on Information Forensics and Security and IEEE Transactions on Communications.



MEHDI BENNIS (Fellow, IEEE) is currently a full tenured Professor with the Centre for Wireless Communications, University of Oulu, Finland, and the Head of the Intelligent COnnectivity and Networks/Systems Group (ICON). He has published more than 200 research papers in international conferences, journals, and book chapters. His main research interests include radio resource management, game theory, and distributed AI in 5G/6G networks. He was a recipient of several prestigious awards. He is an Editor of IEEE Transactions on Communications and the Specialty Chief Editor for Data Science for Communications in the *Frontiers in Communications and Networks*.