IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 January 2023, accepted 2 February 2023, date of publication 6 February 2023, date of current version 14 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243065

SURVEY

Engineering Semantic Communication: A Survey

DYLAN WHEELER, (Graduate Student Member, IEEE), AND BALASUBRAMANIAM NATARAJAN, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA
Corresponding author: Dylan Wheeler (dylan84@ksu.edu)

ABSTRACT As the global demand for data has continued to rise exponentially, some have begun turning to the idea of semantic communication as a means of efficiently meeting this demand. Pushing beyond the boundaries of conventional communication systems, semantic communication focuses on the accurate recovery of the *meaning* conveyed from source to receiver, as opposed to the accurate recovery of transmitted symbols. In this survey, we aim to provide a comprehensive view of the history and current state of semantic communication and the techniques for engineering this higher level of communication. A survey of the current literature reveals four broad approaches to engineering semantic communication. We term the earliest of these approaches classical semantic information, which seeks to extend information-theoretic results to include semantic information. A second approach makes use of knowledge graphs to achieve semantic communication, and a third utilizes the power of modern deep learning techniques to facilitate this communication. The fourth approach focuses on the significance of information, rather than its meaning, to achieve efficient, goal-oriented communication. We discuss each of these four approaches and their corresponding studies in detail, and provide some challenges and opportunities that pertain to each approach. Finally, we introduce a novel approach to semantic communication, which we term context-based semantic communication. Inspired by the way in which humans naturally communicate with one another, this context-based approach provides a general, optimization-based design framework for semantic communication systems. Together, this survey provides a useful guide for the design and implementation of semantic communication systems.

INDEX TERMS 6G, beyond-5G, semantic communication, semantic information theory.

I. INTRODUCTION

While 5G continues to roll out across the globe, the world of wireless communications continues to expand and grow. According to a report published by Ericsson in November of 2021, the monthly global data traffic is predicted to grow exponentially over the next five years [1] (see Figure 1). Recently, the circumstances imposed by the ongoing COVID-19 pandemic have sparked a movement of an increasing number of people choosing to telecommute for work [2]. This will no doubt accelerate global traffic growth even further.

This unprecedented growth is accompanied by an array of new use cases for wireless networks. As defined by the 3rd Generation Partnership Project (3GPP), the 5G network is focused on supporting three main use cases, namely (1) enhanced mobile broadband (eMBB), (2) massive machine-type communication (mMTC), (3) and ultra-reliable low-latency communication (URLLC) [3]. eMBB is

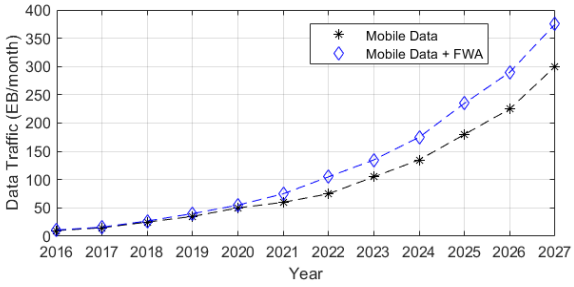

FIGURE 1. Global monthly traffic predictions from mobile data and fixed wireless access (FWA) [1].

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

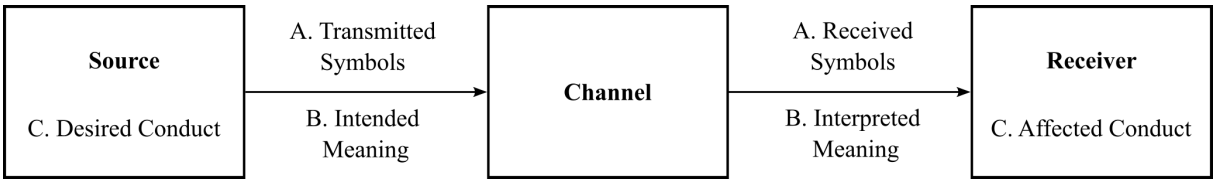

FIGURE 2. General model of communication with components of the three fundamental communication problems.

aimed at providing enhanced services to traditional users of the mobile network, focusing on increased throughput and connection density. mMTC is a paradigm that is designed for a large network of devices each transmitting a relatively small amount of data, e.g. a large wireless sensor network. URLLC is needed when communications are critical and extremely time-sensitive, e.g. when performing remote surgery, or during complex manufacturing processes.

From these observations, we anticipate two trends. First, global data traffic will continue to increase at an exponential rate. As all data communication requires some amount of energy for transmission, this will translate to an exponential growth in overall power consumption of wireless networks. In a world where we all are under increasing pressure to reduce consumption and create more sustainable infrastructure, this trend presents a grand challenge indeed. Second, the decision by the 3GPP to define three specific use cases of the 5G network points to another trend, which is the growing heterogeneity of the wireless network. Communication across the network is increasingly diverse, with a wide variety of use cases and application-specific requirements and constraints. To be able to meet the growing demand for data in a sustainable way, the crucial question is this: *how can we communicate more efficiently over an increasingly heterogeneous network?*

Considering the recent, widespread success demonstrated by artificial intelligence (AI) and machine learning (ML), we believe that a promising approach to address this challenging question is to develop more intelligent communication systems, and specifically, *semantic communication* systems. In their pioneering work, Shannon and Weaver defined three fundamental communication problems [4]:

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

The general model of a communication system is given in Figure 2, with these three problems superimposed. All communication systems today operate at the technical level, i.e., trying to recover transmitted symbols at the receiver as accurately as possible, with no regard for what the symbols mean. Shannon himself stated in his seminal 1948 paper that "the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning…These semantic aspects of communication are irrelevant to the engineering problem" [5].

Clearly, Shannon's view on communication was visionary, and it enabled the extraordinary growth that we have seen in communication systems to this day. However, when subscribing to this view today, we are limited in our options of how to address the previously discussed challenge. If we operate solely at the technical level, an increased demand translates directly to an increased consumption of resources, in the form of either power and/or bandwidth. Increased power consumption is exactly what we are attempting to avoid, and bandwidth is increasingly scarce. Existing usable bandwidths are extremely crowded [6], and there are known issues when operating at higher frequencies (high attenuation, variability, etc.). Recent advances in powerful technologies such as beam forming and massive MIMO [7] can serve as a temporary solution to this problem, but in the face of exponential demand growth, even these will eventually fall short.

Instead of engaging in the unsustainable pursuit of increasing resource consumption to meet demands, some have suggested a turn toward operating at the second level of communication instead, namely the semantic level [8]. If such a communication system is achievable, could it enable more efficient communication? Intuitively, the rationale is sound. To illustrate, let us consider two scenarios of human-to-human communication. In both scenarios, the speaker is trying to teach the listener how compute the area of a circle. In scenario 1, let's say that the listener is someone at least vaguely familiar with geometric concepts, while in scenario 2, the listener is a young child. In both scenarios, the semantic problem is exactly the same; namely, convey what it means to compute the area of a circle. However, the technical problems will likely be very different. In scenario 1, communication may well be very efficient, and perhaps all that is needed is to provide the formula $A = \pi r^2$. In scenario 2, much more information will be required from a technical perspective. To enable understanding of the formula, the speaker would first need to clarify what each piece means, e.g., r represents the radius of the circle, which is the distance from the center of the circle to the edge of the circle, and so on. The speaker would likely need to speak slower and repeat some key points to fully convey the meaning.

What influences the speaker's approach in either scenario is the difference in prior knowledge bases of the listeners, and the *speaker's knowledge* of this difference. Note that, as a trivial solution, the speaker can use the approach in

scenario 2 in both cases, i.e., the speaker can always assume the worst case scenario (no prior knowledge base) and thoroughly explain every aspect of the problem. Let's call this the semantics-agnostic approach. While this will ensure complete conveyance of meaning in both scenarios, it is clear that the speaker is wasting resources (time, energy) by speaking to the listener in the first scenario as if they were a child. This key intuition is the driving force behind the push toward semantically-aware communication systems.

We envision semantic communication as a paradigm shift for future communication systems that embodies a natural progression based on the three communication problems outlined by Shannon and Weaver. By using semantic systems to communicate intelligently, we believe that this technology can meet the challenge posed by rising data demands. Despite its early beginnings in 1952 with Carnap and Bar-Hillel's work on a theory of semantic information [9], the body of literature regarding semantic communications is quite small. Therefore, we feel that a paper surveying the works in this field will be invaluable as the field develops out of infancy. In this paper, we attempt to provide as clear a picture as possible of the current state of semantic communications. Next, we discuss some of the recent works providing their own summaries and visions of semantic communications. We then offer a brief discussion of what it means to define semantics, followed by a presentation of different approaches found in the literature, and finish with our own take on a natural approach to the problem of engineering[1] semantic communication.

A. RELATED WORK

There have been several published works that attempt to provide a vision of what semantic communication might look like, and even fewer that attempt to survey the young field. While most of these have come about in recent years, one of the earliest examples was published in 1992 by Ouksel and Naiman [11], in which they discuss a semantic communication protocol in heterogeneous database systems. They argue that a semantic communication protocol provides a more flexible vehicle of communication and can support effective conflict resolution. Although they refer to this protocol as semantic communication, their ideas align much more with the concept of the *semantic web*, which was first introduced by Berners-Lee et al. [12]. While related, the idea of the semantic web is different from what we refer to as semantic communication. The primary goal of the semantic web is to make the information contained in pages on the internet machine-interpretable. This field has been well-studied over the years since its inception. The semantic web can be thought of as an attempt to achieve the second (semantic) level of communication between the web (source) and a machine (receiver), and thus represents a particular case of general semantic communication. In this paper, we aim to keep our discussion centered on this more general problem.

One of the earliest works providing a vision for addressing the semantic communication problem was given by Rodoplu and Vadvalkar [13]. They introduce their idea of a semantic domain, which includes semantic ''atoms'' and corresponding atomic operations that act on the atoms to generate objects. They then provide their vision of how one might characterize semantic information, making the important observation that the same object might have different semantic information measures in different domains. The vision provided in [13] leans on the idea of using knowledge graphs to represent prior knowledge bases at both the transmitter and receiver, falling within the realm of what we refer to as knowledge graph-based semantic communication, which is the focus of section IV.

Another vision article from 2013 [14] focuses specifically on the problem of semantic misunderstanding. An illustrative example given is the failure of the Mars Climate Orbiter, in which two collaborating teams of engineers were working in different unit systems (imperial and metric), leading to misunderstandings and failure of the mission. The article focuses more on the effectiveness problem rather than the semantic problem, and argues that the key theoretical notion for successful communication is the presence of *sensing*; which is described as feedback to the source indicating successful communication. The vision presented in this article describes agents in a communication system that can learn to achieve successful communication (in the effectiveness sense) despite some initial knowledge mismatch between source and receiver.

In [15], a brief history is provided on the quantification and transmission of information and intelligence. In addition to providing an informative summary on the history of communication techniques and theory, the article raises some important challenges for the design of future intelligent communication, namely:

- Can the formulation of channel capacity include a function of significance?
- How do we define error in the transmission of intelligence?
- How can the code set and signal shaping be defined to support optimal transmission of intelligence?
- How can the receiver be design to optimally accomplish reception of intelligence?

By interpreting the phrase ''transmission of intelligence'' as the nearly synonymous term of semantic communication, these represent some of the grand challenges of practically achieving this higher level of communication. One possible avenue is proposed to address these challenges, which suggests the use of Bayes' decision theory to quantify the ''significance'' of information.

Building on the idea of information significance, in [16] Uysal et al. envision semantics to mean just that: the significance of information as opposed to its meaning. They argue for an extensive cross-layer optimization of the end-to-end

[1]To avoid confusion, we note that by ''engineering,'' we refer to ''the action of working artfully to bring something about.'' [10] Most of the works described in this survey entail theoretical developments, rather than physical systems.

communication system, in a self-described *radical departure* from the well-established way of assessing communication networks. The key idea is that this optimization will yield semantic communication by ''the provisioning of the right piece of information to the right point of computation (or actuation) at the right point in time.'' The development of semantic measures is called for to quantify what information is ''right'' such that the system can be properly optimized. The significance-focused view on semantics is also advocated for in [17], which describes a vision of goal-oriented semantic communication, essentially combining the semantic and effectiveness problems. In addition to outlining this vision, utilizing these ideas is shown to greatly reduce robotic actuation error in a provided example. This view on semantics has much in common with our proposed vision in Section VII, and is discussed in further detail in later sections.

A recent article outlining nine challenges in AI for 6G communications [18] points out the capabilities of recent learning techniques as a potential enabler of semantic communication. The article outlines two challenges which must be met to develop semantic communications, the first of which is the mathematical foundation of semantic communications; while some attempts have been made at defining a mathematical framework on which to build semantic communication, we are still lacking a comprehensive theory. The second challenge is the structure of semantic communication systems, which is posed as a problem of choosing between a general deep neural network (DNN) or further exploring other structural levels of communication to facilitate semantic communication. Another vision article promoting the use of ML in semantic communication for 6G networks is [19]. This extensive article provides a complete view on the current vision for 6G networks, while also discussing details of both semantic and goal-oriented (effective) communications, with various examples and applications given for each. They then provide their vision of the 6G network as incorporating online learning-based communication and control. In contrast to our goal of providing a clear and comprehensive view of the field of semantic communications, [19] focuses on ML and applications of semantic communication. Similar to [18], [19], and [20] provides an overview of end-to-end semantic communications based on DL. The discussion is broken into semantic communications for different modalities, such as text, image and speech. Use cases, including internet of things (IoT) networks and smart factories are discussed, and open issues are presented.

The work that is perhaps the most similar to that presented throughout this survey is the recent review published by Lan et al. [21], in which the authors review principles of semantic communication, discuss existing system architecture designs, and provide an overview of designing semantic communication systems based on knowledge graphs. Their discussion is divided into the categories of human-human (H2H), human-machine (H2M), and machine-machine (M2M) communication, with example applications provided for each. While sharing many similarities, [21] differs from the work here in a few ways. First, the discussion contained in [21] is application-centric, while we aim to focus more on the general techniques and foundations of semantic communication. This leads to a natural contrast in presentation; [21] breaks the discussion into H2H, H2M, and M2M based on the *application* of semantic communication, while we partition topics in our discussion based on the *definition* of semantics in communication. We believe this provides a clearer view of the way in which semantic communications are thought about today. Furthermore, unlike [21], we discuss the classical approaches to quantify semantic information, in order to provide perspective on how the field has evolved since its inception.

As mentioned above, the purpose of this survey is to provide a clear picture of the history and current state of semantic communications, with the hope that it will be a useful guide to those wishing to pursue research within this exciting field. To that end, the rest of this paper will be organized as follows. In Section II, we offer some perspective on the difficulties of defining semantic communication. There have been many different ideas on just how to do this, and we attempt to group these in a natural way. Based on this grouping, Sections III-VI will review works that fall into each category. In Section III, an overview of classical semantic information theory is provided. Section IV reviews works falling under the category of knowledge graph-based semantic communication, which has traditionally been the most common approach. ML-based semantic communication is considered in Section V. This approach has seen a surge of activity in recent years, and is a promising approach moving forward. Section VI goes into more depth on the recently proposed approach of treating semantics as the significance of information. Building off of this idea, in section VII we present our vision on an alternate approach to semantic communication, which emphasizes context as the core component. Section VIII concludes the paper and offers some future research directions.

II. OVERVIEW OF SEMANTICS IN COMMUNICATION

The ambiguity of the word ''semantics'' brings with it an inherent difficulty when attempting to provide a definition. Indeed, this is an issue that has drawn the attention of engineers and philosophers alike. In [22], a brief discussion regarding the philosophical context of semantic communication is provided. There, it is noted that the idea that ''communication must be considered as a means to an end'' was first brought about by Dewey [23], and later ''brought to the forefront of philosophy'' by Wittgenstein [24]. In this work, Wittgenstein fills a short book, organized into a continuous flow of philosophical remarks, with his thoughts and reflections on the fundamental aspects of language. Clearly then, having been at the center of a great amount of philosophical thought, the definition of semantics is a complex one.

However, as we are interested in the engineering of semantic communication, this definition is vital. Without it, we are left blind when attempting to create systems which may

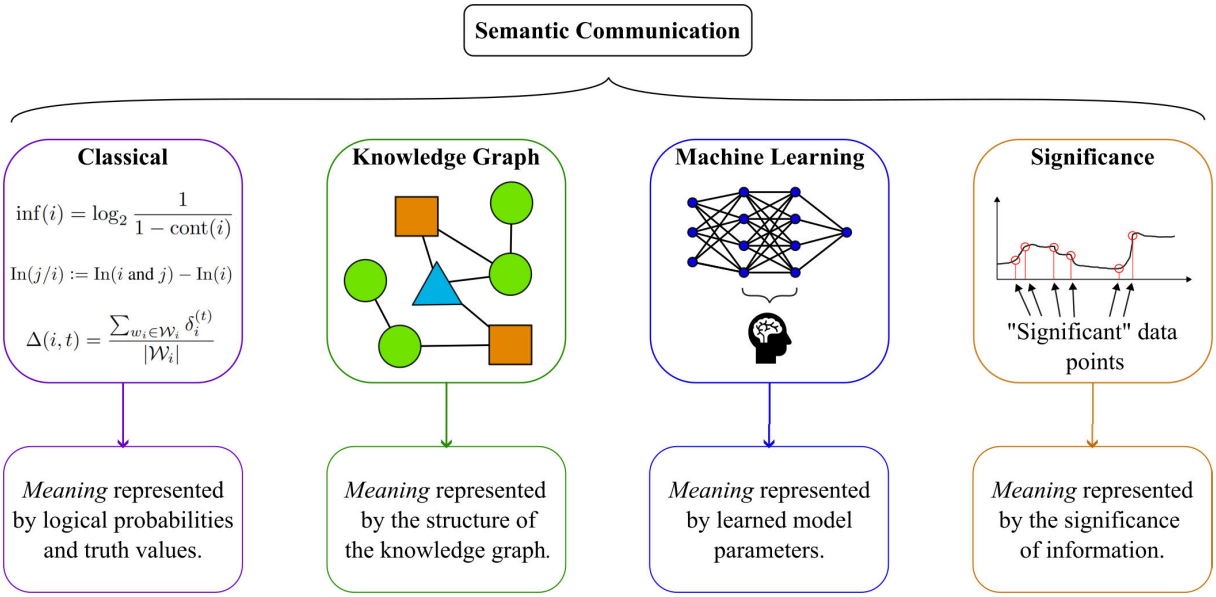

FIGURE 3. The four existing approaches to semantic communication.

achieve this higher level of communication. The solution to any engineering problem requires full understanding of the problem itself, including any simplifications, assumptions, and constraints posed by the problem. Therefore, we dedicate this section to a brief introduction to the different engineering problems that have been posed when attempting to create semantic communication. Inherent in each is the way in which this abstract idea of semantics is defined, which leads to different implications and solutions. Figure 3 provides a high-level view of these approaches and the ways that meaning is represented by each.

As previously mentioned, in their 1952 paper Carnap and Bar-Hillel attempt to outline a Theory of Semantic Communication [9] as a direct response to Shannon's then-recently published "A Mathematical Theory of Communication." The core of their work is to base information measures around *logical* probabilities rather than the *statistical* probabilities which underlie what we now call Information Theory. This definition of semantics is concerned with the so-called logical truth of a statement, from which information measures can be derived. We classify this definition and its derivatives *classical semantic information*. However, it is noted in [9] that this definition of "semantic information is a concept more readily applicable to psychological and other investigations than its communicational counterpart." Regardless, there are still some important works that embrace this idea of semantic information for the engineering problem.

Perhaps the most pervasive method of defining semantics throughout the literature is to do so by using some sort of structured knowledge base. This structured knowledge base can take on many names, such as "semantic network" [25], "taxonomy" [26], "ontology" [27], and others. All of these essentially refer to the same idea, which is to use a graph structure, or *knowledge graph* to represent knowledge in the system [28]. Hence, we refer to these techniques as *knowledge graph-based semantic communication*. Having close ties to the semantic web, it is clear why this approach is popular. By defining knowledge over a graph, it is relatively straightforward to define metrics of "semantic similarity," which can then be analyzed using well-developed graph theory techniques. Furthermore, recent work on graph neural networks brings about the opportunity to incorporate modern learning techniques over such graphs [29].

Another prevalent approach that is seeing a surge of interest is the idea of using ML techniques to "learn" the semantics of a problem. Akin to model-based vs. data-driven approaches to general inference problems (see [30] for more on this), predefined knowledge graphs impose model-based semantics on the problem at hand while ML methods use data to determine these semantics. Borrowing techniques from natural language processing (NLP) and computer vision, deep networks can be taught to communicate in the most efficient manner while preserving semantic content [31]. Similarly, by implementing reinforcement learning (RL) methods, these networks can be refined over time and even adapt to dynamic changes in the communication problem. We refer to this approach as *machine learning-based semantic communication*. Initial studies examining this approach to semantic communication over several modalities have emerged in recent years [30], [31], [32], [33].

Finally, a fourth definition which differs significantly from those previously described is the one first mentioned in Section I-A: the idea of semantics as the "significance" of information; we call this *significance-based semantic communication*. While classified as an approach toward semantic communication, this approach essentially addresses the third

TABLE 1. Summary of works in classical semantic information.

Theory of Weakly Semantic Information	Carnap & Bar-Hillel [9]	Semantic information using logical probabilities
	Bao *et al.* [38]	Definition of semantic entropy, corresponding theorems
	Basu *et al.* [39]	Semantic compression
Theory of Strongly Semantic Information	Floridi [36]	Extension of TWSI using truth values
Truthlikeness	D'Alfonso [37]	Semantic information based on truthlikeness

level of communication problem: the *effectiveness* problem. Rather than concern ourselves with the meaning of a message, advocates of this approach call for communication of the *right* information. Of course, what information is "right" will depend on the application and the desired outcome, therefore leading to the idea of *effective* or *goal-oriented communication*. This approach lends itself well to machine-machine communication, in which we are less concerned with the conveyance of "meaning" and more concerned with what the system accomplishes. One well-studied way of quantifying what information is "right" is the popular age of information metric [35], which relates to generating and delivering information at the right time. By defining other metrics to quantify what is "right" for the problem at hand, a joint optimization can be carried out to achieve optimal communication.

These four approaches to defining semantics roughly partition the previous research regarding semantic communication. Based on this analysis of the current literature, we broadly define semantics as *any definition of information or the transfer thereof that considers something beyond the statistical nature of the symbols used to represent that information.* This view unites the definitions discussed above, despite their differences in the "something beyond" that is considered by each. Of course, these definitions are not mutually exclusive. ML can be used to determine which information is "right," just as classical semantic information metrics can be used to design and tune neural networks. Just as model-based deep learning (DL) incorporates both prior knowledge and data-driven techniques for inference, graph neural networks can be used to learn knowledge graphs for semantic communication. However, by treating each of these definitions individually, a complete picture is given of the current state of semantic communication. In the next sections, we dive into each definition, the engineering approaches that come as a result, and some of their potentials and challenges.

III. CLASSICAL SEMANTIC INFORMATION

As mentioned earlier, Carnap and Bar-Hillel's paper *An Outline of a Theory of Semantic Information* attempts to tackle the problem of engineering semantic communication through the quantification of semantic information [9]. This quantification is the first step towards efficient semantic communication, as it allows us to consider ideas such as semantic compression and semantic error. Therefore, we begin our discussion with the ideas and methods related to classical semantic information. The works discussed in this section are summarized in Table 1.

A. THEORY OF WEAKLY SEMANTIC INFORMATION

Following the convention of [36], we will refer to the theory laid out by Carnap and Bar-Hillel as the Theory of Weakly Semantic Information (TWSI), for reasons that will be discussed later. TWSI is defined over a language system $\mathcal{L}_n^k$ which contains n "things" (or individuals) and k primitive one-place predicates (descriptors). An *atomic* sentence is said to consist of a single predicate describing a single thing, while a *molecular* sentence is formed from two or more atomic sentences joined with some logical connective, including: or, and, if…then, if and only if. Any sentence can either be logically true, logically false, or logically indeterminate. Furthermore, *logical relations* are defined. For example, for sentences i and j, we have i *logically implies* j defined to mean that "if i then j" is logically true. A *state description* is a sentence in which each of the k predicates is specified for each of the n individuals; thus completely specifying all aspects of the universe. Common set notation can be used to talk about "classes" of entities within the universe. For example, [9] describes a system consisting of three individuals, $\{a, b, c\}$, and two binary predicates, young or old (Y or O) and male or female (M or F). Then an example of a state description could be given as "(a is F and Y) and (b is M and Y) and (c is M and O)." Some other possible states are given below in Table 2.

TABLE 2. Some states of an example universe.

a		b		c	
M/F	Y/O	M/F	Y/O	M/F	Y/O
F	Y	M	Y	M	O
F	O	M	O	M	O
M	Y	M	Y	M	Y
F	O	M	Y	F	Y

Note that, for binary predicates, a universe will consist of 2^{nk} possible state descriptions. Similar to the process in which Shannon developed entropy as a measure of information [5], Carnap and Bar-Hillel begin with requirements/axioms that

semantic information (and its corresponding measures) must satisfy. Denoting the semantic information of a sentence as In(·), the first axiom is given as

$$\text{In}(i) \text{ includes } \text{In}(j) \iff i \text{ logically-implies } j, \tag{1}$$

meaning that i says everything that is said by j (and possibly more) if and only if j is implied by i within the logical framework. For example, take $i = (a \text{ is } F \text{ and } Y)$ and $j = (a \text{ is } F)$. Clearly, j is implied by i, and thus the semantic information of i includes that of j (though the converse is not true). This axiom requires us to treat information as a *set* or *class* of something; it is important to note that the *amount* of information is arbitrary at the moment, and must be defined on this set. Some theorems are derived from this axiom, and the concept of *relative information* is defined as

$$\text{In}(j/i) := \text{In}(i \text{ and } j) - \text{In}(i) \tag{2}$$

where $\text{In}(j/i)$ is again some set or class. Continuing the example above, we can see that $\text{In}(j/i) = \emptyset$ (the empty set), since $\text{In}(i \text{ and } j) = \text{In}(i)$. However, $\text{In}(i/j) \neq \emptyset$.

Based on (1) and (2), a concept of the information of a sentence is defined, and is termed the *content* of a sentence. Content is derived from what is called a *content-element*, which is simply the negation of a state description. The content of a sentence i, denoted $\text{Cont}(i)$, is then defined as the set of content-elements logically implied by i. Intuitively, $\text{Cont}(i)$ can be thought of as the set of state descriptions in the overall state-space (i.e. universe described by our language system) that are *eliminated* with knowledge of the sentence i. Take $i = (a \text{ is } F \text{ and } Y)$ as before. Out of the $2^6 = 64$ possible state descriptions in the universe, this sentence eliminates all 48 in which a is not F or a is not Y, and $\text{Cont}(i)$ is composed of the negations of all such state descriptions. Clearly then, a self-contradiction will "say the most" (by eliminating all state-descriptions) and a tautology will "say the least" (by eliminating no state-descriptions). Similarly, a complete state description can be thought of as carrying much information, since it eliminates all other state descriptions. Note that this notion of information has nothing to do with the truth of a sentence, which is a point we will revisit. With this idea of information in place, the primary question is addressed: how shall the *amount* of information be defined?

The amount of semantic information carried by sentence i is denoted as $\text{in}(i)$, and the following requirements are stated:

$$\text{Cont}(j) \subseteq \text{Cont}(i) \Rightarrow \text{in}(i) \geq \text{in}(j) \tag{3}$$

$$\text{Cont}(j) \subset \text{Cont}(i) \Rightarrow \text{in}(i) > \text{in}(j) \tag{4}$$

$$\text{Cont}(i) = \emptyset \Rightarrow \text{in}(i) = 0 \tag{5}$$

Note the subtle difference between In(·) and in(·); In(·) represents the information itself, while in(·) is used to quantify *how much* information is given by In(·) Using Cont(·) as the information content of a sentence, these requirements are straightforward and make intuitive sense. By (3) and (4), a sentence containing all the information of another should have a greater than or equal amount of information, with equality only if the information carried is the same. By (5), the amount of information is zero if the information of a sentence is the empty set.

Based on these requirements, two measures are offered as the main contribution of [9]. The first is termed the *content-measure* of a sentence, denoted cont(·) (different from Cont(·)), and is defined as any proper m-function of the negation of a sentence. We will not go into the details of what constitutes an m-function here (see [9, Section 6] for further reading), but suffice it to say that it satisfies (3)-(5), and defines a measure taking values between 0 and 1, thus representing a logical probability measure. However, a problem arises with this measure regarding another intuitive requirement not yet stated, namely additivity. Just as in classic Information Theory, we would like the information measure of two *independent* sentences to follow additivity, i.e. $\text{in}(i \text{ and } j) = \text{in}(i) + \text{in}(j)$ for i, j independent; it is shown that cont(·) does not satisfy this intuition [9, Thm. 6-15]. Thus, a second measure is proposed and is termed *measure of information*, denoted by inf(·), not to be confused with the infimum operator. This second measure is defined as

$$\text{inf}(i) = \log_2 \frac{1}{1 - \text{cont}(i)}. \tag{6}$$

Observe that this measure is analogous to the classical information-theoretic definition of entropy, making use of the logical probability $\text{cont}(i)$ instead of the statistical probability $p(i)$.

A comparison of these two measures is given in [9], and it is shown that both exhibit intuitively desirable properties, and likewise they both exhibit intuitively undesirable properties. The lack of additivity of cont(·) is one example. Another is that inf(·) lacks a counterpart to the property of cont(·) stating that $\text{cont}(i \text{ and } j) \leq \text{cont}(i) + \text{cont}(j)$. Thus, it is concluded that neither represent an ideal measure of semantic information, but rather that they both have specific strengths and weaknesses.

B. THEORY OF STRONGLY SEMANTIC INFORMATION

A problem with TWSI occurs when presented with a sentence that constitutes a contradiction, e.g., "i and not i." As mentioned above, under the definition of Cont(·), this sentence would carry with it maximum semantic information. Intuitively, however, we know that a contradiction *should* carry no information; it is obviously untrue and leaves the receiver no less informed than before the reception of the message. This ambiguity manifests itself in the mathematics of TWSI as well, and is known as the Bar-Hillel-Carnap Paradox (BCP). Therefore, in [36] Luciano Floridi proposes that *truth* lies at the root of this paradox, which can be solved by the incorporation of truthfulness considerations into TWSI.

The theory which follows is outlined in [36] and is deemed a Theory of Strongly Semantic Information (TSSI). Again, the goal is to develop a theory from which semantic information can be quantified, which would clearly be a useful theory

for the engineering of semantic communication. As a starting point, three desiderata are given:

- D.1 Avoid any counterintuitive inequality comparable to BCP.
- D.2 Treat the alethic (truth) of a sentence not as a supervenient but as a necessary feature.
- D.3 Extend a quantitative analysis to the whole family of information-related concepts.

The core of TSSI is the definition of degrees of vacuity and inaccuracy. The intuitive idea is that semantic information is related to "how true" and "how false" a sentence is. A highly vacuous sentence is one that is true, but carries with it little information. Similarly, a highly inaccurate sentence is false and also carries little information. This brings about the strange idea that a false sentence may carry more information than a true sentence. As an example, consider the system of three individuals and two predicates described earlier. The tautology $i =$ "(a is F) or (a is M)" clearly provides no information, yet it is always true nonetheless. In contrast, consider the sentence $j =$ "(a is F) and (b is M) and (c is F) and (a is Y) and (b is Y) and (c is O)" when c is actually M. While j is false, it is not too difficult to reason that it carries more information than i.

Mathematically, these concepts are formalized as a positive or negative degree of "semantic distance" of a sentence i from a fixed point, which is defined as the given situation w to which i is supposed to refer. True statements take on positive degrees between 0 and 1, while false statements take on negative degrees between -1 and 0. This mapping is denoted by the function $f(i)$. For false statements, the *degree of inaccuracy* simply counts the number of false atomic statements e in i and divides by the total number of atomic statements, or the *length* l of i

$$f(i) = -e(i)/l(i), \tag{7}$$

where i is a false sentence. On the other hand, the *degree of vacuity* is more difficult to define since all atomic statements of a true sentence are true. Therefore, this degree is defined as the number of situations n (including the true situation) with which i is consistent divided by the total number of possible situations (s^l for a system with s predicates),

$$f(i) = n(i)/s^l, \tag{8}$$

where i is a true sentence, and a "situation" is nothing more than a complete state description as was defined for TWSI. Note the lack of symmetry between (7) and (8); this is one argument against TSSI.

Using these degrees, the *degree of informativeness* is defined as

$$\imath(i) = 1 - f^2(i). \tag{9}$$

Continuing the previous example, the tautology is consistent with all situations, and thus has a degree of vacuity $f(i) = 2^6/2^6 = 1$ and a degree of informativeness of $\imath(i) = 1 - 1^2 = 0$. Meanwhile the false statement j has degree of inaccuracy $f(j) = -1/6$ and a degree of informativeness $\imath(j) = 1 - (-1/6)^2 \approx 0.972$, and we see that $\imath(\cdot)$ is consistent with intuition. The relationship between degrees of vacuity and inaccuracy and degree of informativeness is given in Figure 4.

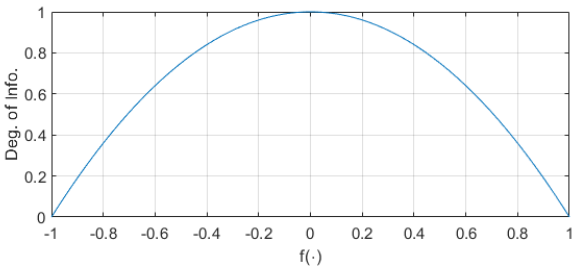

FIGURE 4. Relationship between degrees of inaccuracy, vacuity, and informativeness.

The *amount of vacuous information* is then defined as the integral of this curve from 0 to the degree of vacuity; the maximum amount of vacuous information is then simply the integral from 0 to 1, yielding 2/3. Defining this maximum as α and the amount of vacuous information carried by i as $\beta(i)$, the amount of strongly semantic information carried by i is denoted as $\imath^*$ and is defined as

$$\imath^*(i) = \alpha - \beta(i) \tag{10}$$

It is finally argued that $\imath^*(i)$ provides a solution to the BCP, in that it shows that "semantic information about a situation presents an actual possibility that is inconsistent with at least one but not all other possibilities. A contradiction is not information-rich because it is not a possibility." It is stated that in TSSI, a contradiction is simply a limit instance of "uninformation," or lack of both positive and negative misinformation.

C. SEMANTIC INFORMATION WITH TRUTHLIKENESS

TSSI is a step in the right direction, but still has some shortcomings with regards to quantifying semantic information. First, the degrees of vacuity and inaccuracy are inherently asymmetric; the former is a measure depending on the model, and the second is a measure based solely on the sentence at hand. In addition, it is unclear how to quantify vacuity and inaccuracy for more complex sentences beyond simple conjunctions of atomic statements.

In [37], Simon D'Alfonso builds on the foundations laid by TWSI and TSSI by expanding on Floridi's idea of using "truthlikeness" for quantifying semantic information. Two existing approaches to quantifying truthlikeness are offered. The first, termed the Tichie-Oddie approach, computes truthlikeness, denoted Tr, as the compliment of some distance function $\Delta(\cdot)$ from the statement i and the true statement t [37]:

$$\mathrm{Tr}(i) = 1 - \Delta(i, t) \tag{11}$$

TABLE 3. Summary of classical semantic information measures.

Method	Measure	Benefits	Drawbacks
TWSI [9]	$\text{cont}(i)$	• Simple measure satisfying (3)-(5) • Entirely dependent on logical probabilities	• Does not satisfy independent additivity • Trouble dealing with contradictions (BCP)
	$\inf(i) = \log_2 \frac{1}{1-\text{cont}(i)}$	• Satisfies (3)-(5) and independent additivity • Entirely dependent on logical probabilities	• Does not satisfy triangle inequality • Trouble dealing with contradictions (BCP)
TSSI [36]	$f(i) = -e(i)/l(i)$ (i false), $f(i) = n(i)/s^l$ (i true)	• Avoids BCP by considering truth • Satisfies intuition that true and false statements carry information	• Asymmetric: different measures for true and false statements • Unclear how to define for more complex statements
Truthlikeness [37]	$\text{Tr}(i) = 1 - \frac{\sum_{w_i \in \mathcal{W}_i} \delta_i^{(t)}}{\lvert\mathcal{W}_i\rvert}$	• Avoids BCP by considering truth • Provides flexibility through adjustment of atom weights	• Addition of false statement can increase information yield
	$\text{Tr}(i) = 1 - \gamma\Delta_{\min}(i,t) + \lambda\Delta_{\text{sum}}(i,t)$	• Avoids BCP by considering truth • Satisfies several adequacy conditions	• May not be applicable in certain scenarios

where $\Delta(\cdot)$ takes values in $[0, 1]$. We can see that (10) is a particular case of (11) with $\Delta(i,t) = f^2(i)$. Let $\mathcal{W}_i$ denote the set of states in which statement i is true. Furthermore, define $\delta_i^{(t)}$ as the number of mismatched atomic statements between state w_i and the true state w_t, i.e., the number of false atomic statements in w_i. This value is weighted by the inverse of the number of propositions in the universe. The Tichie-Oddie approach [37] specifies the distance function as

$$\Delta(i,t) = \frac{\sum_{w_i \in \mathcal{W}_i} \delta_i^{(t)}}{|\mathcal{W}_i|} \tag{12}$$

Continuing with the example in the previous subsection, i agrees with all state descriptions, so $|\mathcal{W}_i| = 64$. After computing the sum in the numerator we have $\Delta(i,t) = 32/64 = 1/2$ regardless of the true state. Thus, a statement that is always true provides little information no matter the true state of the universe. For the second involving sentence j, we obtain $\Delta(j,t) = 1/6$ and $\text{Tr}(j) = 5/6$, again matching intuition that the false j carries more information than the true i.

In general, the function $\Delta(\cdot)$ can be any distance measure between the state specified by the sentence i and that of the true statement t. The Niiniluoto approach [37] makes use of a different distance metric, namely the min-sum measure

$$\Delta_{\text{ms}}^{\gamma\lambda}(i,t) = \gamma\Delta_{\min}(i,t) + \lambda\Delta_{\text{sum}}(i,t) \tag{13}$$

where

$$\Delta_{\min}(i,t) = \min_{w_i \in \mathcal{W}_i} \Delta(w_i, w_t) \tag{14}$$

$$\Delta_{\text{sum}}(i,t) = \frac{\sum_{w_b \in \mathcal{W}_i} \Delta(w_i, w_t)}{\sum_{w_b \in B} \Delta(w_b, w_t)} \tag{15}$$

and B is the set of all states in the logical space. The state distance is defined as

$$\Delta(w_i, w_t) = \frac{\delta_i^{(t)}}{n} \tag{16}$$

where n is some atomic weight. Niiniluoto shows that the min-sum measure of (13) satisfies certain adequacy conditions.

Finally, D'Alfonso proposes a novel measure, termed the *value aggregate* measure, which is claimed to lie in between the Tichy/Oddie and Niiniluoto approaches. First, each state is assigned a value, denoted val, with

$$\text{val}(w) = \frac{t^{(w)}}{n2^n} \tag{17}$$

where $t^{(w)}$ is the number of true atoms in state w, and n is the number of propositional variables in the logical space. The following algorithm is used to calculate information yield for a statement i:

1. Determine $\mathcal{W}_i$.
2. Place members of $\mathcal{W}_i$ into an array X_1 and order from lowest to highest value.
3. Create empty array X_2 of length 2^n. Fill the first $|\mathcal{W}_i|$ elements with the array X_1. Use the last (highest value) element of X_1 to fill the remain spaces of X_2.
4. Sum the values of X_2 to get the information measure.

It is claimed without proof that the value aggregate measure satisfies many of the adequacy conditions listed by Niiniluoto. A summary of the measures discussed in this section is provided in Table 3.

D. EXTENSIONS OF TWSI

These approaches to quantifying semantic information, namely TWSI, TSSI, and truthlikeness, all provide the groundwork for a theory of semantic information, which could enable semantic communication. In [38], TWSI is used as a foundation for a general, abstract semantic communication model. First a semantic information source is defined as a tuple (W_s, K_s, I_s, M_s), where

- W_s is the world model (potentially observable by the source),

- K_s is the background knowledge of the source,
- I_s is the inference procedure of the source,
- M_s is the message generator used by the source.

Similarly, the semantic receiver is defined by the analogous tuple (W_r, K_r, I_r, M_r). The source model is further specified by assuming that the world model W_s is a set of interpretations with probability distributions μ; for the familiar example of propositional logic, an interpretation would be a set of positive propositions. The inference procedure I_s is then defined as a satisfiability reasoner for the propositional logic, and the message generator M_s employs a fixed coding strategy. The model entropy is then given by

$$H(W) = -\sum_{w \in W} \mu(w) \log_2 \mu(w) \tag{18}$$

Letting $m(x)$ denote the logical probability of a message x within this model, the *semantic entropy* of x is defined as

$$H_s(x) = -\log_2(m(x)). \tag{19}$$

Note that this is equivalent to the TWSI measure given by (6) with $m(x) = 1 - \text{cont}(x)$. When the knowledge base K_s is included in this formulation, the set of possible worlds is restricted to a set compatible with K_s. This brings about the notion of *conditional semantic entropy*, where $m(x|K)$ now denotes the logical probability of x conditioned upon the background knowledge, and we have

$$H_s(x|K) = -\log_2(m(x|K)). \tag{20}$$

Letting X be a finite set of allowed messages with probability distribution $P(X)$, we know the classic Shannon entropy of X as

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x). \tag{21}$$

The following theorem relates the model (semantic) entropy (18) to the message (syntactic) entropy (21):

Theorem 1: $H(X) = H(W) + H(X|W) - H(W|X)$.

Proof: See [38]. □

The main implication of Theorem 1 is a formal method of quantifying semantic uncertainty that is rooted in the TWSI and relating it to classic Shannon entropy. As entropy is used to quantify information in classic Information Theory, this gives a way of comparing the syntactic information and the semantic information under the present world model.

The authors in [38] then discuss the idea of using semantics for data compression. Intuitively, the idea is that some messages may be semantically equivalent without being syntactically equivalent. For example, many times we are able to understand the true meaning of text despite some minor errors in spelling or grammar. Hence, a syntactic error does not necessarily induce a semantic error. Based on this principle, the idea is that we can achieve maximum compression by choosing the smallest semantically equivalent message for communication. Let $\overline{X}$ denote the smallest subset of X such that each $x \in X$ is semantically equivalent to some $\overline{x} \in \overline{X}$.

Theorem 2: For a semantic source with interface language X, there exists a coding strategy to generate a semantically equivalent interface language X' with message entropy $H(X') \geq H(\overline{X})$. No such X' exists with message entropy $H(X) < H(\overline{X})$.

Proof: See [38]. □

Theorem 2 provides bounds on the maximum achievable data compression give a model as described above. Note that Theorem 2 is analogous to the classical source coding theorem, in that it gives existence of such a code but no insight into how to design the coding strategy.

The final major result of [38] is the so-called Semantic Channel Coding Theorem. Some notations used include:

- $I(X;Y) = H(X) - H(X|Y)$ is the traditional information-theoretic *mutual information* between X and Y
- $\overline{H_s(Y)} = -\sum_y p(y) H_s(y)$ is the average logical information of received messages

Theorem 3: For every discrete memoryless channel, the channel capacity

$$C_s = \sup_{P(X|W)} \{I(X;Y) - H(W|X) + \overline{H_s(Y)}\} \tag{22}$$

has the following property: For any $\epsilon > 0$ and $R < C_s$, there is a block coding strategy such that the maximal probability of semantic error is $< \epsilon$.

Proof: See [38]. □

Similar to Theorem 2, Theorem 3 provides a result that parallels the classic Channel Coding Theorem of Information Theory. This result gives a bound on the maximum amount of information that can be transmitted for some arbitrary probability of semantic error.

In [39], the general model and results of [38] are extended and practical semantic compression algorithms are given based on graph theoretic results. However, basic definitions, such as the semantic source and receiver, are different from those in [38], making it difficult to relate this work to previous results. Perhaps the most useful contribution of [39] is a discussion of some practical techniques for semantic compression.

The first suggested idea is to allow *non-uniquely decodable* codes. This fits the case in which a certain syntactic message represents two semantically equivalent states, i.e., message x could be decoded as either state a or b. Another idea is to extend the concept of erasure channel codes, such that some bits are intentionally "erased." With this approach, only partial information may be recovered at the receiver, with some intentional semantic ambiguity.

A practical algorithm is then suggested for a system in which the source and the receiver share a knowledge base that is defined as a bipartite fact-conclusion graph, see Figure 5. The problem studied is that in which the source wants to convey a set of conclusions to the receiver in as few symbols as possible (where both facts and conclusions can be transmitted and are equally expensive). This problem reduces to computing the *minimum-vertex cover* which is solvable in

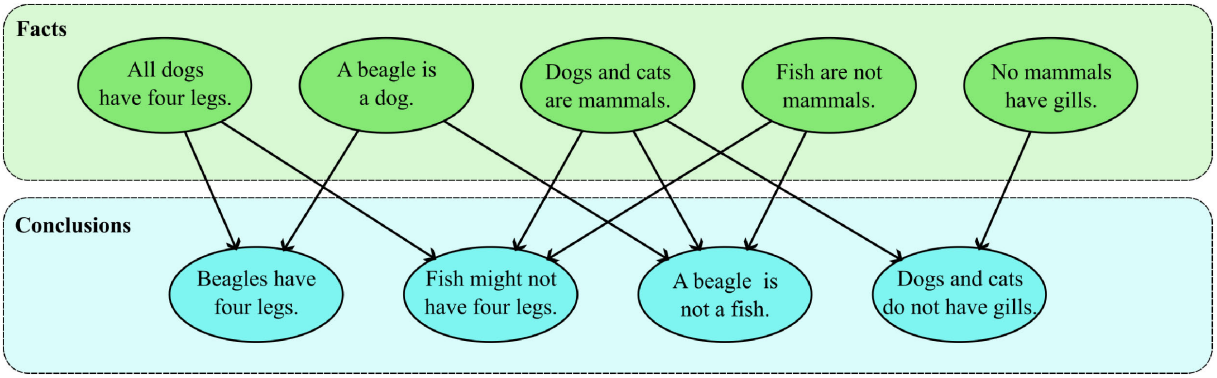

FIGURE 5. Example bipartite fact-conclusion graph model of a knowledge base.

polynomial time for bipartite graphs. This simple example represents a situation in which semantic compression is computationally feasible, given the assumption on background knowledge.

E. SUMMARY

The classical semantic information approach attempts to mirror the path followed in the monumental development of Information Theory to define and quantify semantic information. TWSI attempts to achieve this through the use of logical probabilities rather than statistical probabilities. However, this theory contains a paradox in which a contradiction carries maximal semantic information. As a remedy, TSSI introduces the use of truth values to quantify semantic information, based on the idea that both false and true statements can carry varying degrees of information. This idea has been extended through the introduction of various truthlikeness measures for quantifying semantic information. Recently, some have attempted to extend the original TWSI, relating semantic entropy to traditional information-theoretic entropy and providing analogous source-coding and channel-coding theorems. All of these approaches seek to quantify semantic information of a sentence through the use of logical probabilities and truth values.

With regards to the motivating problem described in Section I, namely, the trend of exponentially increasing global data traffic, the works described in this section provide no quantitative results addressing this issue. This is due to the fact that these works are more concerned with the development of semantic information theory, rather than the subsequent use of this theory for semantic communication. That is not to say that the presumed benefits are nonexistent, however, and future work in this area should seek to confirm this potential.

F. CHALLENGES AND OPPORTUNITIES

Given the ability to quantify semantic information, development of optimal coding techniques (source and channel) should follow. However, as can be seen from the previous discussion, this quantification is no trivial task. A definition of semantic information itself is elusive, and attempts to quantify it become mired with paradoxes and counter-intuitive results. However, should such a complete theory exist, the benefits that would follow are clear, thus presenting a major opportunity.

Previous work seems to point to the idea of truthlikeness as the best approach to quantifying semantic information. The main challenge with this approach is the need for a large knowledge base. Simple examples are given using propositional logic models consisting a few objects and propositions, but the size of these models explodes for more realistic and practical systems. Furthermore, these models must be predefined and known at both the source and receiver, which introduces further complications. Another challenge is selection of the correct information measure. As we have seen, there is no single measure for semantic information (yet); the available measures all come with their respective strengths and shortcomings.

A clear opportunity lies in the further development of the theory of semantic information. While significant progress has been made on the subject, a unifying and ubiquitous theory does not yet exist. Another promising opportunity is the incorporation of existing semantic information measures into practical systems. In particular, *neurosymbolic AI* has become a field of great interest as of late [40]. By combining the strengths of symbolic logic and DL, neurosymbolic AI could enable powerful learning systems that are capable of logical reasoning. By applying neurosymbolic AI to the problem of semantic communication, perhaps the logical model at the source and receiver could be learned from data. Then, the DL architecture could be used to efficiently facilitate semantic communication. As all of the previous work on semantic information quantification relies to some degree on propositional logic, the application of neurosymbolic AI seems to be a perfect fit.

IV. KNOWLEDGE GRAPH-BASED SEMANTIC COMMUNICATION

Semantic communication necessarily requires a knowledge base at both the transmitter and receiver. The engineering of

TABLE 4. Summary of works in knowledge graph-based semantic communication.

Knowledge Representation	Delgado-Frias & Moore [25]	"Semantic network" as a knowledge graph
	Mertoguno & Lin [27]	Evolution of a distributed knowledge base
	Swartout *et al.* [41]	Using ontologies to share knowledge
Semantic Similarity	Rada *et al.* [26]	Link-based similarity measure on knowledge graphs
	Resnik [43]	Node-based similarity measure on knowledge graphs
	Jiang & Conrath [44]	Similarity measure considering both links and nodes
	Sathya & Uthayan [45]	Measure of quality for an entire ontology
Semantic Sensor Web	Sheth *et al.* [46]	Semantic sensor web
	Gyrard *et al.* [49]	Semantic sensor web for machine-to-machine communication
	Chun *et al.* [50]	IoT directory for the semantic sensor web
	Bhajantri & Pundalik [51]	Data-processing for the semantic sensor web
	Schachinger & Kastner [52]	Semantic interface for building automation
	Lakka *et al.* [53]	Interoperability of semantic systems
Semantic Communication	Jeong *et al.* [54]	Semantic error correction of spoken queries
	Zhang *et al.* [55]	Model for semantic natural language processing
	Li [48]	Text analysis and character recognition with a KG and ML
	Wang *et al.* [58]	Hybrid KG/ML model for explanation of recommendations
	Aumayr *et al.* [59]	Recommendation system for wireless network operation
	He *et al.* [60]	Model for general goal-based communication
	Güler *et al.* [61]	Optimal semantic communication with game theory
Working with Knowledge Graphs	Wei *et al.* [62]	Reasoning over large knowledge graphs
	Zheng *et al.* [63]	Embedding of large-scale knowledge graphs
	Zhu *et al.* [64]	Fusion of knowledge graphs

semantic communication thus requires some form of knowledge representation to encapsulate the knowledge at the transmitter and receiver. One prevalent method of representing knowledge is through the use of a *knowledge graph* (KG). Although a particular KG may be defined in different ways, in general a KG can be said to use a graph structure to model a given knowledge base. For example, nodes of the graph could represent objects, while edges represent relations between the objects. By referencing the KG, a system can then perform communication that is semantically-aware. The works discussed in this section are summarized in Table 4.

A. KNOWLEDGE REPRESENTATION WITH KGs

One of the earliest works making use of this idea was published in 1989, which proposes a so-called "semantic network architecture" for AI processing [25]. Arguing that knowledge representation and manipulation is required for artificial intelligence, a multi-processor architecture is proposed. This semantic network takes the form of a KG, in which nodes are defined as *concepts* and edges are defined as *relationships* between concepts. Figure 6 gives a simple example of a KG. The architecture is composed of a grid of processing elements (PEs) which compute the corresponding links and nodes associated with the semantic network. By building the architecture in such a manner, it is argued that the system can perform intelligent actions, such as labelling the scene of a given image (computer vision).

Other early studies into knowledge representation with KGs include [27] and [41], published in 1996 and 1997, respectively. In [27], the problem of a *distributed* knowledge base is studied using KGs. Specifically, the challenges of evolving the distributed knowledge and controlling this evolution are considered. This distributed knowledge base is framed as a multi-agent system in which some communication exists between the agents. Within the system, each agent possesses a simple KG and agents collaborate to perform global inferencing. Participating agents receive distributed rewards after correct inferences, resembling a mutli-agent reinforcement learning scheme. A crossover operator is adopted from genetic algorithms to facilitate knowledge evolution. Distributed knowledge bases are also the focus of [41], which uses the idea of a large-scale shared ontology. Here, an ontology is defined as "a hierarchically structured set of terms for describing a domain," from which a knowledge base can be constructed. The key idea is that if two knowledge bases are formed from the same ontology, knowledge can

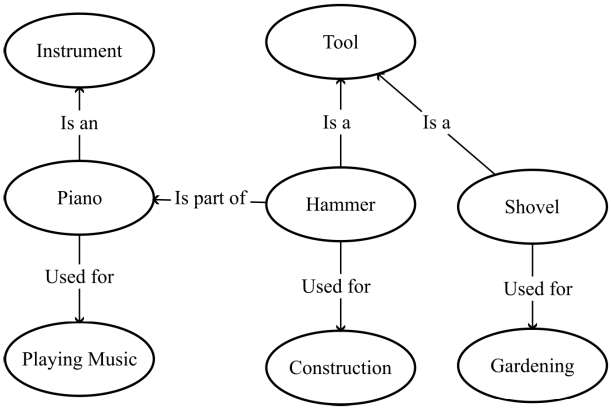

FIGURE 6. An example knowledge graph.

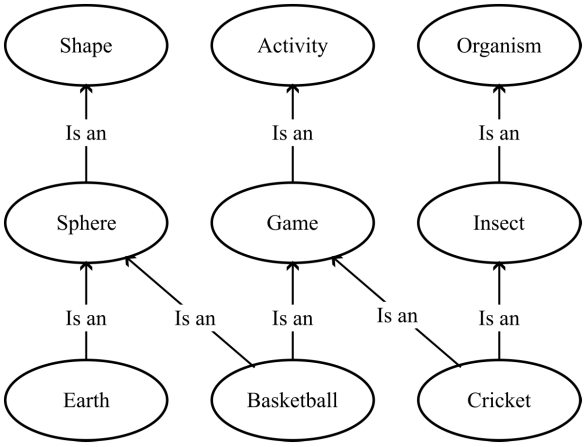

FIGURE 7. Example of an *is-a* KG, where all edges represent an *is-a* relationship.

be easily shared between the two. An analogy would be two people from the same culture; the shared cultural norms and ideas would serve as the so-called ontology, and though their knowledge bases will not be the exact same, intuitively it will be easier for them to communicate given their shared background. A set of desiderata for ontologies is given, essentially proposing that ontologies should represent large-scale, living documents from which we can define smaller, application-specific knowledge bases. This idea is the basic premise behind the Web Ontology Language (OWL) standards that shape the semantic web [42].

B. SEMANTIC SIMILARITY MEASURES ON KGs

As was seen in the previous section, an important idea for semantic communication is the idea of semantic similarity. In the sense of classical semantic information, logical probabilities and truthlikeness were used to define this similarity. How should one quantify semantic similarity given a KG representation? This is a question that some began to address around the time KGs emerged as a way of representing knowledge. In [26], a metric of semantic similarity, or distance, is proposed and simply termed Distance. The authors assume a KG that consists of an *is-a* hierarchy, see Figure 7 below. This Distance metric is defined for *sets* of concepts and is dependent upon the path lengths between nodes in a KG; formally, it is defined as the average minimum path length over all pairwise combinations of nodes between two sets of nodes, i.e., for sets of nodes $\mathcal{X}$ and $\mathcal{Y}$,

$$\text{Distance}(\mathcal{X}, \mathcal{Y}) = \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} d(x, y), \tag{23}$$

where $d(x, y)$ is the shortest path between x and y. For example, consider the distance between the sets {Sphere, Earth} and {Basketball} in the KG shown in Figure 7. The shortest path between Sphere and Basketball is 1, while the shortest path between Earth and Basketball is 2, yielding a total distance of 3/2. Repeating for the sets {Sphere, Earth} and {Cricket}, we obtain a distance of 7/2, and thus we can conclude that the concepts {Sphere, Earth} are more similar to the concept {Basketball} than {Cricket}. It is shown through experimental results that (1) Distance can simulate human assessments of conceptual distance and (2) Distance can evaluate some cognitive aspects of semantic networks. However, it is also found that Distance is less applicable to nonhierarchical KGs.

An alternate measure is presented in [43], again for an is-a taxonomy KG. First, it is noted that link-based measures, such as (23), suffer from the fact that links in the taxonomy are assumed to represent uniform distances, while in reality some linked concepts may be "closer" than others. Indeed, for the average person in the United States, {Basketball} probably shares a stronger intuitive link to {Game} than {Cricket} does, while this association may differ elsewhere. Therefore, a *node-based* measure is proposed based on the notion of information content. First, the KG is augmented with a function $p : \mathcal{C} \to [0, 1]$ where $\mathcal{C}$ is the set of all nodes in the graph. $p(c)$ can be thought of as the *probability* of encountering a concept c; thus, concepts higher in the taxonomy will have greater probability. Then we can define the semantic similarity between two concepts as

$$\text{sim}(c_1, c_2) = \max_{c \in S(c_1, c_2)} -\log p(c) \tag{24}$$

where $S(c_1, c_2)$ is the set of concepts that subsume *both* c_1 and c_2. Intuitively, this measure computes the log-inverse probability of the *most-specific* node (farthest "down" in the taxonomy) which branches to both concepts; therefore, the farther down this node, the smaller its probability, and the greater the similarity measure. For example, in Figure 7, this would imply that {Cricket} is more similar to {Basketball} than to {Earth}.

Combining (23) and (24), [44] develops a new measure for the is-a taxonomy KG that provides an even higher correlation with human similarity judgement benchmarks. It involves a link-based calculation which takes into consideration node-based edge weights. First, information content is defined in the same way as [43]:

$$\text{IC}(c) = -\log p(c). \tag{25}$$

TABLE 5. Summary of KG-based semantic similarity measures.

Method	Measure	Benefits	Drawbacks
Link-Based [26]	$\text{Distance}(\mathcal{X},\mathcal{Y}) = \frac{1}{\lvert\mathcal{X}\rvert\lvert\mathcal{Y}\rvert}\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} d(x,y)$	• Simple to compute • Intuitive	• Each link assumed as equi-distant • No node-based information
Node-Based [43]	$\text{sim}(c_1,c_2) = \max_{c\in S(c_1,c_2)} -\log p(c)$	• Links not necessarily equi-distant • Interpretable, based on probabilities	• Requires additional function p
Node/Link-Based [44]	$\text{Dist}(c_1,c_2) = \min_{\text{path}(c_1,c_2)}\sum_c \text{wt}(c,p_c)$	• Incorporates both ideas • Achieves highest human correlation	• Requires additional function p • Complicated to compute

Then it is argued that the *strength* of a child link is dependent on the information content of the parent node,

$$\text{LS}(c,p_c) = \text{IC}(c) - \text{IC}(p_c), \tag{26}$$

where p_c is the parent node of c. This link strength, among other factors, is used to compute an overall edge weight $\text{wt}(p_c,c)$ between the child c and parent p_c. Finally, the semantic similarity, denoted as Dist, is defined as summation of edge weights along the shortest path linking two nodes

$$\text{Dist}(c_1,c_2) = \min_{\text{path}(c_1,c_2)} \sum_{c\in\{\text{path}(c_1,c_2)-\text{LSuper}(c_1,c_2)\}} \text{wt}(c,p_c) \tag{27}$$

where $\text{LSuper}(c_1,c_2)$ is the lowest super-ordinate of c_1 and c_2. Note that in [44], $\text{Dist}(w_1,w_2)$ is used, where w_1 and w_2 are introduced to address the scenario when one node belongs to multiple inheritances. An evaluation of this metric shows a correlation value of 0.828 with human similarity judgements, higher than the node-based and link-based measures alone. This implies that considering link strength within a KG can lead to much more reasonable similarity judgements. A summary of the KG-based semantic similarity metrics described here are given in Table 5.

While (23), (24), and (27) all give ways to measure semantic similarity within a KG, a more recent work [45] proposes a semantic metric to assess the quality of an entire ontology. It is argued that most existing metrics for ontology assessment consider only structural properties, and ignore the semantics of the ontology. The proposed metric is termed the ''relationship deviation metric'' and is determined by the number of breadthwise and depthwise relationships in the KG. It is shown that the proposed metric captures the quality of some example ontologies.

C. THE SEMANTIC SENSOR WEB

One prevalent use of KGs for semantics is the so-called semantic sensor web (SSW) proposed in [46], which uses ideas from the semantic web. Namely, metadata is captured along with the desired data from each sensor, such that better sense can be made of the observations (similar to how metadata can be used to determine what is contained within a webpage). For example, providing the total lifetime of a sensor may lead to more informed decision-making; a sensor that has been operating well past its expected lifetime may be more likely to produce faulty measurements. With regards to communication, this metadata can provide the *context* necessary for semantically efficient communication within the network. The suggested core set of attributes in [46], as adopted from the RDFa language [47], are

- *about*: a triple that specifies the resource metadata is about
- *rel* and *rev*: specify a relationship or reverse-relationship with another resource
- *href*, *src*, and *resource*: specifies the partner resource
- *property*: specifies a property for the content of an element
- *instanceof*: optional, specifies the RDF type of the object

Furthermore, [46] advocates for the use of ontologies along the three types of semantics associated with sensor data: spatial, temporal, and thematic. Once these ontologies have been defined, rule-based reasoning can be implemented to provide better inferences from sensor observations.

Building on this idea of an SSW, [49] proposes a semantic-based approach to automatically combine, enrich, and reason about machine-to-machine (M2M) data to support IoT applications. One key idea is that the *meaning* of new information is pre-defined in an ontology, and therefore the ontology can facilitate the fusing of cross-domain knowledge. Suppose we possess an ontology which stores weather-related knowledge and soil-related knowledge and the relations between the two; by leveraging this shared knowledge, it might be possible to achieve smarter fusion of the data for agricultural decision-making. A concept called ''Linked Open Rules'' is defined as a means of sharing and reusing semantic rules, and some examples are given that demonstrate the proposed concept, including a weather monitoring application. In [50], an IoT directory, called IoT-DS, is proposed to support semantic description, discovery, and integration of new objects as an alternative approach to building a SSW. One key difference is that IoT-DS distinguishes static and dynamic components, based on whether other attributes vary with time. It is shown that IoT-DS provides a 40% reduction in communication overhead as compared to a naive approach.

Some more recent works that look at the idea of a SSW are [51], [52], [53]. In [51], a survey of data processing

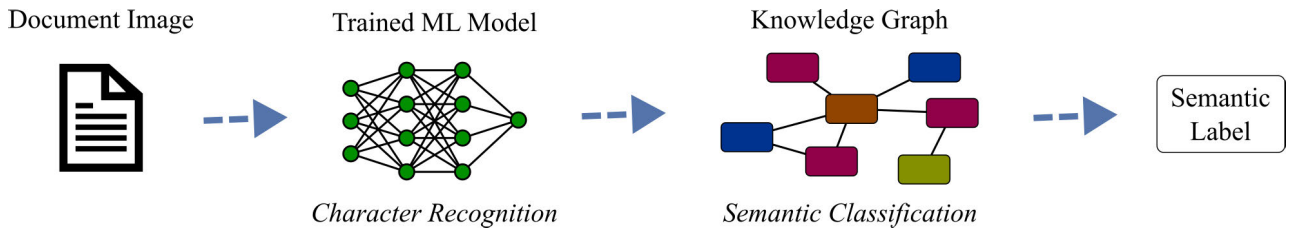

FIGURE 8. Simplified diagram of the framework proposed in [48] for semantic document analysis.

techniques for SSW is provided, as well as another take on the architecture of a SSW, which is partitioned into physical, semantic, application and controller layers. Reference [52] considers again the problem of a semantic interface for M2M communication, this time with the intended application of building automation. Requirements are defined for the building automation problem, from which an interface is developed and an ontology formed. Reference [52] provides a great example of how KG-based semantics can be utilized to engineer a SSW for a specific application. Finally, [53] studies the general problem of semantic interoperability, or the ability to interact and exchange data with shared meaning, between systems. An interoperability mechanism termed SEMIoTICS is proposed, in which an IoT application request is received by a directory which then connects the corresponding sensor and actuator to fufill the request.

D. KNOWLEDGE GRAPHS FOR SEMANTIC COMMUNICATION

The works discussed in the previous subsection are similar to the idea of the semantic web, in that they focus more on how to semantically describe objects rather than the task of semantic communication itself. One of the first works utilizing a KG for semantic communication is [54], which focuses on semantic error correction for spoken query processing. Spoken query processing, or question answering, has become a hot topic as of late (more on this in Section V). Reference [54] proposes the use of two KGs: a *domain* dictionary and an *ontology* dictionary. The first represents application-specific knowledge, while the second contains the pure general knowledge of the world. In the agricultural setting mentioned above, the ontology dictionary might contain general information that does not vary between crops or location, while the domain dictionary might consist of site-specific information. For a spoken query, the semantic recovery stage involves the use of a semantic confusion table based on the domain knowledge to replace semantic errors. Lexical, or syntactic, recovery is then performed based on the corrected semantic phrase. Experiments performed on the domain of in-vehicle telematics show that the technique yields a 37% reduction in term-error-rate as compared to baseline models. This decreased error rate bodes well for traffic reduction in communication systems, as less information will need to be retransmitted due to errors.

In recent years, some have sought to apply KG-based semantics to text-based communication. Reference [55] looks to apply KGs to the problem of natural language processing (NLP), and proposes an enhanced language representation model termed ERNIE, which is an enhancement of the popular NLP model BERT [56]. The model operates by first recognizing entities in some text, and matching these to entities in a pre-defined KG. The KG representation is then embedded using known algorithms, such as TransE [57], and then used in conjunction with standard text embeddings as the input to an aggregator. DL encoder-decoder techniques are then used to perform common NLP tasks. In [48], a hybrid KG-ML approach is proposed to perform text analysis through character recognition. The model first uses DL to perform character recognition over two bodies of text, then uses semantic measures to quantify how similar the bodies of text are. To quantify the semantic similarity, each word is modeled as a word in a graph, and corresponding distance metrics are proposed. The general framework is illustrated in Figure 8. Through experiments carried out on Dickinson's Portfolio, it is concluded that the performance of the proposed technique can meet real-time recognition requirements.

Another application of KGs for semantic communication involves recommendation systems, in which the goal is for an automated system to make the best possible recommendations to some user. One way that KGs have been used for this task is to enhance explainability of the recommendations [58]. Explainability can enhance a user's experience when receiving recommendations, e.g., Amazon suggesting products to a user "based on previous purchases." In [58], a model termed *Knowledge Path Recurrent Network* (KPRN) is proposed as a hybrid KG-ML model. This model works on a KG which contains objects in the recommendation system and performs reasoning based on the paths in the KG to infer user preference. In the Amazon example, products would be modeled as nodes in the KG, and the software might suggest to a user a product with the shortest path length to that just purchased by the user. A long short-term memory (LSTM) network is adopted to model the sequential dependencies of objects and relations, from which a pooling operation is used to obtain the prediction. By modeling the sequential dependencies, the system can offer an explanation for each prediction. One scenario where a recommendation system is useful is wireless network management, where an automated system can recommend the best course of action to an

operator. The growing complexities of these networks can make management difficult for a human operator, while automated methods might not always make the best decisions. Thus, a hybrid approach has some benefits. Reference [59] focuses on gathering context from a wireless network and correlating it with useful information from documents in the network provider's domain using KGs. The KG is formed from two types of documents: product troubleshooting manuals and incident/troubleshooting reports by technicians. The first provides well-structured problem solving instructions, while the second provides important links between symptoms and issues. This KG is then used to reason about new problems that arise, and recommend a course of action for the operator. Experiments show a decrease of up to 91% of the documents that are presented to an operator, drastically reducing the amount of information involved in the communication process.

Finally, there has been some work on the more general scenario where two agents strategically communicate to achieve some goal, in which KGs are used to facilitate communication. Reference [60] studies a symmetric collaborative dialogue setting in which two agents communicate to achieve a common goal. Each agent possesses private knowledge. The model, termed Dynamic Knowledge Graph Network (DynoNet) models the dialogue state as a KG which evolves as the conversation advances. The graph contains three types of nodes, namely item, attribute, and entity nodes. These nodes are embedded, and used as an input to a LSTM network. Provided this embedding and an embedded utterance from another agent, the network generates an utterance in response. Experiments show that DynoNet is able to hold a coherent and strategic conversation with a human, and that the number of entities and attributes uttered to achieve the goal are reduced by 47% and 17%, respectively, when compared to baseline rule-based communication strategies. Goal-oriented communication naturally lends itself to game-theoretic analysis, which is the focus of [61]. In goal-oriented communication, as in game theory, there are two or more agents that seek to achieve some goal, and can employ multiple strategies to reach said goals. Similar to [60], in [61] communication is modeled as taking place between two agents, with the addition of a third agent who could aim to either improve/deteriorate communication performance. The optimal transmission policies are characterized, where optimality is defined as minimizing end-to-end semantic error. This error is derived from the semantic similarity measures proposed in [26], [43], and [44]. The interaction is modeled as a Bayesian game, where uncertainty is introduced about the characteristics of other agents. It is shown that, in the static scenario, that finding encoding/decoding strategies to minimize average semantic error is an NP-hard problem, and two algorithms are proposed. In addition, it is demonstrated that when the third agents signals its true nature to the communicating agents, a sequential equilibrium is attainable, i.e., when sufficient information is available regarding the intentions of agents involved in the communication, efficient semantic communication can be achieved. It is shown that judicious transmission policies can greatly reduce semantic errors.

E. WORKING WITH KGs

In the previous subsections, we have seen how KGs can be used to facilitate semantic communication. If our goal is to engineer semantic communication through the use of KGs, then we must develop effective methods of working with KGs. Reasoning over, i.e., deriving knowledge from, a KG is a challenge that becomes more difficult with increasing scale of the KG. In [62], a reasoning system is proposed for large-scale KGs. This system, termed KGRL, is based on the web ontology language 2 rules logic (OWL2 RL) which was developed for the semantic web. Using the rules defined by OWL2 RL, the iterations of the reasoning procedure are reduced based on dependency relations and multiple applications of these rules. Experiments show that KGRL is able to greatly increase reasoning efficiency as compared to state-of-the-art reasoning systems.

Many of the works previously discussed combined KG methods with DL methods, which requires an embedding of the KG into some vector space. This embedding essentially aims to preserve the knowledge represented by the graph in the embedding space. Similar to reasoning, this task becomes difficult at large scales. Reference [63] studies the problem of training KGs at scale, proposing a technique termed DGL-KE to efficiently perform KG embeddings. DGL-KE provides optimized embeddings for three types of hardware configurations: (1) many-core CPU machines, (2) multi-GPU machines, and (3) a cluster of CPU/GPU machines. For each hardware type, the DGL-KE takes advantage of parallel processing to fully utilize the computing hardware. The allocation of memory resources throughout the process is specifically designed for each hardware type, and mini-batch training is utilized to perform the embedding. Other optimization techniques employed by DGL-KE are graph partitioning, negative sampling, data access to relation embeddings, and applying gradients to global embeddings. Experiments for hardware types (1) and (2) demonstrate improved efficiency compared to other methods.

Another interesting problem is that of KG fusion. For example, say that when a new agent joins the network, we wish for its knowledge to be merged with the overall knowledge of the network. If both knowledge bases are represented by KGs, how should we fuse them together? One way is by *instance matching*, which establishes a semantic link between instances in KGs. Reference [64] proposes a method called Follow-the-Regular-Leader Instance Matching (FTRLIM), which is able to match instances between large-scale KGs with approximately linear time complexity. The FTRLIM framework is based on an blocking algorithm called MultiObJ, which divides instances into blocks and is also developed in [64]. Through various experiments, FTRLIM is shown to perform effective and scalable KG fusion.

For more information on the field of KGs, we direct the interested reader to [28], which provides a recent and comprehensive survey covering (1) KG representation learning, (2) knowledge acquisition and completion, (3) temporal KGs, and (4) knowledge-aware applications (such as semantic communication).

F. SUMMARY

Following the large amount of research that has been dedicated toward the semantic web, there have been many works which seeks to utilize KGs for semantic communication. This idea stems from the fact that some kind of knowledge representation is required for semantic communication, and a common way of representing knowledge is with KGs. Furthermore, some have proposed similarity measures for concepts within a KG, which can be used to quantify semantic similarity. This form of knowledge representation has been proposed for use in the so-called semantic sensor web, which extends the semantic web to physical sensor networks. Various works have attempted to leverage KGs for different semantic communication scenarios, including recommendation systems and general goal-oriented communication. Methods such as hybrid ML-KG systems, game theoretic techniques, and others have been applied to achieve this communication. Finally, some have studied the specific problem of working with the KG itself, which will be important when implementing these systems at scale. The key idea behind KG-based semantic communication is that knowledge, and therefore meaning, can be captured by a KG and utilized by the semantic system.

Some of the works mentioned above have provided quantitative results demonstrating improved performance with regards to communication efficiency. As mentioned, results in [50] indicate a reduction of around 40% in overall network traffic. Results from [54] show decreased error rates using semantic techniques, which in turn has implications for reduced communication traffic. Experiments in [60] show that goal-oriented semantic communication can reduce the amount of entities and attributes communicated by 47% and 17% respectively. In a more specific example, results from [59] demonstrate that semantic communication can reduce the amount of total information conveyed in a recommendation system by up to 91%. These results, stemming from diverse examples and use-cases, show that KG-based semantic communication can address the issue of increasing data demands with more efficient communication.

G. CHALLENGES AND OPPORTUNITIES

KGs are a popular way of representing knowledge in a system, which can then be used to facilitate semantic communication. However, this approach does not come without its challenges. First of all, as knowledge in the system grows, the KG can become massive. As was discussed in the previous subsection, working with KGs becomes difficult as they grow larger. Scalability is a challenge that must be addressed for efficient semantic communication. This difficulty is amplified in systems with stringent communication requirements, such as those requiring real-time operation. Another challenge is that a KG must be predefined with some prior knowledge of the system, reducing the ease of deployment. Building these graphs can be time-consuming and therefore costly.

These challenges present opportunities as well. Further development of scalable methods, such as those in [62], [63], [64], will be required for the practical use of large (and thus more expressive) KGs. Furthermore, techniques such as transfer learning present a promising approach toward reusing existing KGs for new applications, such that a KG does not need to be built from scratch for each deployment. Another exciting field is of ML and DL on graphs. These techniques can be employed to learn optimal embeddings and relations from existing data, and bring with them all of the benefits that have been achieved with DL in other domains.

V. MACHINE LEARNING-BASED SEMANTIC COMMUNICATION

The field of ML has seen an explosion of activity in recent years. At its core, ML seeks to learn from data in order to better perform some task. The availability of massive amounts of data and advanced algorithms have enabled the practical implementation of ML in many domains, including NLP [65], computer vision [66], and others. Over the past few years, some have sought to utilize the power and flexibility of ML to develop semantic communication systems. The general idea behind these approaches is to ''learn'' the semantics of the problem. Just as in image processing, where the key features for classifying an image may be hidden to us, the semantic aspects of communication may be unknown. Through the use of ML methods, these latent semantic features can be learned through training and added to the system's knowledge base automatically. Thus, by utilizing learning methods, these systems address one of the key challenges inherent to KG-based semantic communication, namely the requirement of a predefined knowledge base. In this section, we will take a closer look at some of these works to illustrate the ML-based approach to semantic communication. The works discussed in this section are summarized in Table 6.

A. DEEP LEARNING METHODS

DL is a subset of ML which utilizes *deep neural networks* (DNNs) to perform prediction and decision-making tasks [67]. These networks are trained through an iterative update of the network parameters, which is typically accomplished through some gradient-based method. Modern DNNs come in many different forms, such as the classic multi-layer perceptron [68], convolutional neural network (CNN) [69], recurrent neural network (RNN) [70], etc. As with any DL problem, a key consideration is how to choose the *loss function*, which will determine how the parameters are tuned. For more on information on DL, please see [67].

Regarding semantic communication, DL can be used in many different ways, often depending on the modality of

TABLE 6. Summary of works in machine learning-based semantic communication.

Deep Learning	Lu *et al.* [71]	Ensuring mutual understanding using a LSTM-RNN
	Hua & Du [72]	GAN for cross-modal retrieval
	Huang *et al.* [73]	Semantic coding of images with a GAN
	Qiao *et al.* [76]	CNN, RNN for scene text recognition
	Tong *et al.* [78]	Audio-based semantic communication with CNN and federated learning
	Zhou *et al.* [79]	Text-based semantic communication using a transformer
	Sana & Strinati [81]	End-to-end communication with semantic symbols using a transformer
	Xie *et al.* [31]	Text-based semantic communication with DeepSC
	Weng *et al.* [33]	Speech-based semantic communication with DeepSC-S
	Xie *et al.* [34]	Visual question answering with MU-DeepSC
	Xie & Qin [32]	Semantic communication for IoT with L-DeepSC
Reinforcement Learning	Lu *et al.* [87]	General semantic communication with arbitrary similarity function
	Wang *et al.* [89]	Reinforcement learning over a knowledge graph for text communication
	Lotfi *et al.* [90]	Collaborative deep reinforcement learning with heterogenous agents
	Yun *et al.* [91]	Reinforcement learning for air-to-ground semantic communication

communication (text, images, speech, etc.). One of the first works employing DL for semantic communication provides an illustrative example. In [71], aviation radiotelephony communication (ARC) is considered, where a statement spoken by one party is repeated by the other to ensure understanding. In this scenario, clear and reliable communication is crucial, as misunderstandings can lead to accidents and crashes. To minimize semantic errors (i.e., misunderstandings between pilots and air traffic control), a long short-term memory RNN (LSTM-RNN) is proposed. The LSTM-RNN takes as its input a pilot-air traffic control sentence pair. To train the network, statement pairs are assigned a similarity R, ranging from $R = 0$ to $R = 1$, with 1 indicating the strongest semantic similarity. Clearly, a strong similarity is desired; we want the pilot and air traffic control to be on the same page. The network is then trained to minimize the *cross-entropy* error:

$$L = -\sum_{n=1}^{N} R^* \log(R) + (1 - R^*) \log(1 - R) \qquad (28)$$

where R^* is the true label value and R is the predicted consistency value. After training, the network can be used to classify new statement pairs; if a pair is found to be semantically inconsistent, a signal can be given to either party to correct the misunderstanding.

Another interesting problem involving semantic communication is that of cross-modal retrieval, in which the format of the query is different than that of the information being queried, e.g., a voice request for certain textual document in a database. In [72], a generative adversarial network (GAN) is proposed to address this problem by performing *semantic correlation* on multi-modal data, specifically image and text data. A GAN is a network composed of two networks,

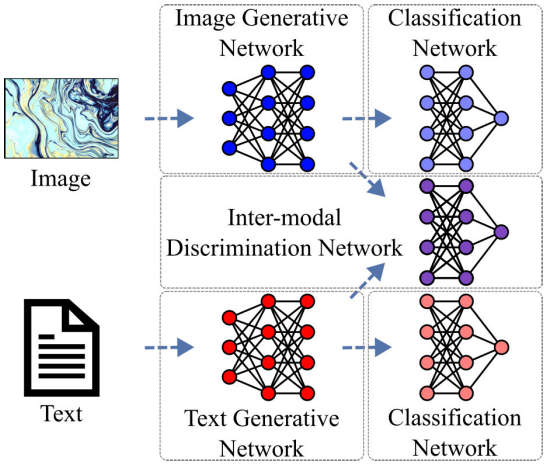

FIGURE 9. Basic framework of the multi-modal GAN approach proposed in [72].

namely a generator and a discriminator. The generator learns by attempting to "fool" the discriminator, while the discriminator is optimized to discern between outputs of the generator. [72] proposes a framework in which the generator takes both an image and a body of text as inputs and learns their representations. The goal of the discriminator is to distinguish between the two modalities. The basic framework of this approach is illustrated in Figure 9. After training, the generator will have learned the representations of the heterogeneous modalities in a common space, in which data with similar semantics will be close, i.e., the *meaning* common to both the text and the image is captured in this space. The proposed method is shown to outperform both traditional and deep methods with respect to the mean average precision metric.

Another work utilizing a GAN is [73], in which DL is implemented with the goal of semantic coding of images. The goal of semantic coding is to minimize the bit rate of transmission while preserving the semantic information of the image. Again, the semantics of interest here are unknown, and the goal is to learn them throughout the training process. Here, the generator network is used to learn and restore semantic information which is used as a "base layer" of the image. The generator loss function is formed as a rate-perception-distortion trade-off, including a combination of the VGG loss [74] and LPIPS [75]. Then, Better Portable Graphics (BPG) residual coding is used to refine the image. The overall network explores different strategies to optimize the rate-perception-distortion tradeoff, and is shown to exhibit similar performance to baselines models while utilizing a 2-4 times reduced bit rate.

Similar to [55] and [76] also looks to enhance standard encoder-decoder NLP techniques (this time without the use of a KG), specifically for the problem of recognizing some text within an image, or *scene text recognition*. For example, in the case of a self-driving vehicle, it would be beneficial for the vehicle to be able to recognize the text located on street signs within the field of vision, as a human driver would. In their model, termed SEED, semantic information is used in both the encoder module for supervision and the decoder module for initialization. The semantic information is predicted from the image features which are first extracted with a CNN and RNN. To predict this information, a simple fully-connected DNN is trained with a dual cross-entropy and cosine embedding loss function. Then, the image features and the semantic information are both fed to the decoder to perform text recognition. Using ASTER [77] as an exemplar to demonstrate the framework, it is shown that the semantic enhancement improves performance of the model.

A CNN along with federated learning (FL) is proposed to facilitate audio-based semantic communication in [78]. FL is a branch of DL in which distributed, local models are trained individually and then combined to form a global model. In [78], FL is implemented in a system consisting of a single server and many devices. The devices train local models based on data, to reduce communication overhead to the server. Each local model consists of an autoencoder with a convolutional layer for extracting the semantic information of the audio. Normalized root mean square error is used as a loss function to evaluate the quality of semantic reconstruction, and experiments show that the proposed architecture can achieve around 100 times improved performance (with respect to the mean square error) with around 1/3 the transmitted data of traditional methods.

Yet another type of DNN is the *transformer*, which has seen wide success in the field of NLP [65]. Transformers rely on the idea of *attention*, which provides different weights to different features of the input data, similar to how our brains pay more attention to certain perceptual inputs. In [79], an adaptive universal transformer is implemented for text-based semantic communication. The optimization goal is "to minimize the semantic errors while facing different communication situations." The universal transformer is able to accomplish this by adding a circulation mechanism which can dynamically allocate greater computation time to semantically complex statements. Intuitively, this is similar to how a human might read some text. Passages with a simple meaning are easier to understand and thus can be read faster, while passages with complex meaning demand more thought (equivalently, computation time). Cross-entropy is used as the loss function for training the network, and simulations performed on the standard proceedings of the European Parliament [80] show an improved performance over traditional methods.

The transformer has also been proposed for semantic communication in [81]. In this work, an end-to-end architecture is proposed which performs communication with *semantic symbols*, which are used to represent semantics. The semantic communication model is defined, including the encoder, decoder, channel and noise. The transformer model is then designed for this model, including both source and channel coding in a joint architecture. This architecture is trained using the cross-entropy loss, and experiments in a NLP setting demonstrate that the proposed system can achieve a similar performance to traditional techniques with a 21% decrease in the number of symbols.

1) DeepSC AND ITS VARIANTS

One particular DL model for semantic communication that has been the subject of various studies was first proposed in [31], where it was given the name DeepSC. DeepSC is also based on the transformer DL model, and the model proposed in [79] takes its inspiration from DeepSC. Similar to other works, [31] defines the model for the semantic communication system, which consists of a transmitter performing both semantic encoding and channel coding, and a receiver performing the inverse operations. The architecture of the proposed network is shown in Figure 10. Both transmitter and receiver possess some background knowledge. The goal is stated as simultaneously minimizing semantic errors (measured with the cross-entropy loss function) and transmitted symbols. This is accomplished through an end-to-end transformer network, which uses a self-attention mechanisms for extracting semantic information from text; here, the meaning is captured by this attention mechanism and which input text is emphasized by the model. Various metrics are used to demonstrate the superior performance of DeepSC compared with traditional communication methods.

There have been a few variants inspired by DeepSC, which all aim to facilitate semantic communication within different modalities. One example is [33], which proposes a semantic communication system for speech signals, termed DeepSC-S. It is claimed that the end-to-end communication system "learns and extracts the essential speech information." This is a very intuitive idea, as it is clear how non-verbal qualities of speech can impact a conversation (tone, volume, etc.).

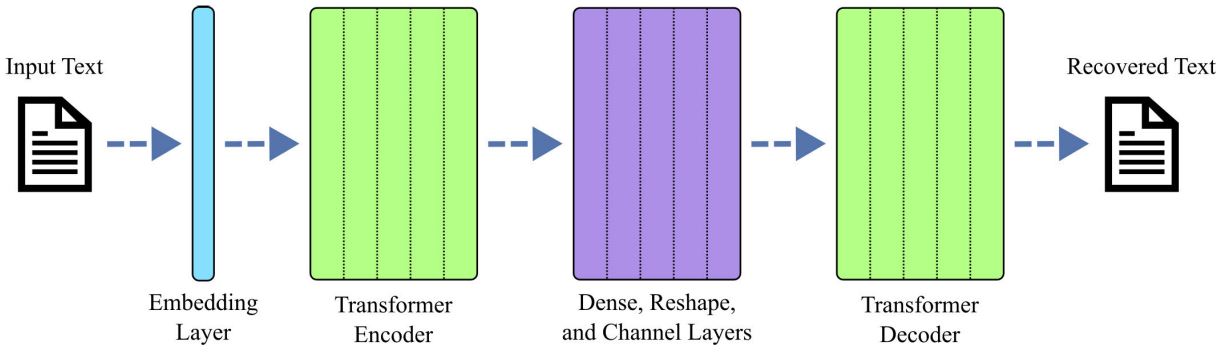

FIGURE 10. Architecture of the end-to-end semantic communication system DeepSC proposed in [31]. Each box represents a different section of the network, which are all trained in a joint manner.

Specifically, DeepSC-S employs an attention-based CNN for speech coding and a CNN for channel coding, and the mean-squared error loss function is used for training. It is confirmed through simulations that DeepSC-S outperforms traditional systems with equivalent bit rates under AWGN, Rayleigh and Rician channels.

Another variation of DeepSC is proposed in [34], which considers a multimodal communication system and is termed MU-DeepSC. The task-oriented system is implemented for visual question answering (VQA), where the transmitter sends an image and a question about that image (text) and the receiver aims to correctly answer the question (e.g., is there a red ball present?). The transmitter consists of two networks; one employs ResNet-101 [82] and a CNN for semantic image encoding, and the other utilizes Bi-LSTM [83] and a dense DNN for semantic text encoding. These encodings are transmitted over the channel, where the receiver then implements a memory, attention, and composition (MAC) network to generate the answer. The end-to-end network is trained with the cross-entropy loss, and simulation results using the CLEVR dataset [84] demonstrate accuracy gains of up to 80% compared to traditional methods with a 70% reduction in transmitted symbols.

As we've seen in the discussion of the semantic sensor web, a prevalent idea is the use of semantic communication in IoT applications. In [32], a "lite" version of DeepSC, named L-DeepSC, is proposed for use with IoT networks. In L-DeepSC, the semantic communication model is trained and updated in the cloud and distributed to IoT devices. These devices then implement the model to perform semantically aware data collection and text transmission with low complexity. The system first uses a least-squares estimator to obtain channel state information (CSI), and then a deep de-noise network to refine the CSI estimates. The trained model is then compressed through sparsification and quantization and broadcast to the IoT devices. The devices then use this model to perform semantic communication with text data, uploading new data to the cloud. It is shown that the proposed system performs competitively with traditional methods, especially in the low SNR domain. Moreover, L-DeepSC reduces the model parameters of the original DeepSC network by around 60%, which translates to reduced communication upon model distribution to IoT devices.

As ML-based semantic communication is based on *learning* the semantics of the problem, explicit semantic metrics are not defined. However, the chosen *loss function* affects how the semantics are learned. Table 7 provides the loss functions used in each of the discussed DL approaches.

B. REINFORCEMENT LEARNING METHODS

Another popular approach to ML is reinforcement learning (RL) [85]. In reinforcement learning, the model is viewed as an agent in a state space. From its current state, the agent can take some action, for which it is provided a reward. The goal of the agent is to discover the action-taking policy which will maximize long-term rewards from any given state. Rather than use existing data to determine this policy, in RL the model is trained by letting the agent "explore" different policies; the model takes some sequence of actions, and then tunes the parameters based on the rewards received. A simple illustration of this learning process is given in Figure 11. Due to the agent/reward set up of RL, it is a method which is well-suited for learning to play different games, such as chess and Go. Indeed, learning models including both RL and DL

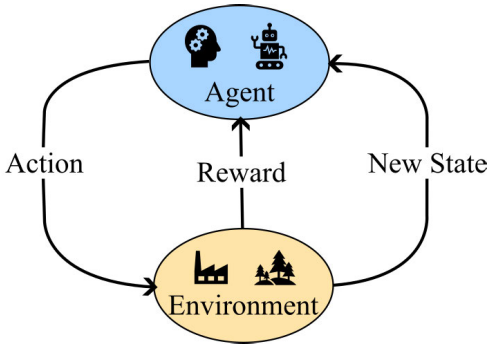

FIGURE 11. Reinforcement learning framework.

TABLE 7. Summary of DL-based semantic communication loss functions.

Loss Function	Expression	Description
Cross-entropy [31], [32], [34], [71], [79], [81]	$L = -\sum_{n=1}^{N} R^* \log(R) + (1 - R^*) \log(1 - R)$	Promotes semantic similarity or probability values R consistent with labels R^*
Cross-modal generative [72]	$L_{gen} = \lambda L_{adv} + L_{cls}$, $L_{adv} = -\sum_{i=1}^{N} [\log(1 - D(G_I(x_i))) + \log D(G_T(y_i))]$, $L_{cls} = -\sum_{i=1}^{N} \log \left[C(G_I(x_i))^T c_i + \beta C(G_T(y_i))^T c_i\right]$	Aims to promote similar image embeddings $G_I(x_i)$ and text embeddgins $G_T(y_i)$ given semantically similar inputs x_i and y_i
Rate-perception-distortion [73]	$\mathcal{L}_{E,G} = \mathcal{L}_G + \lambda_1(\mathbb{E}[d(x, G(\hat{w}, s))] + \mathbb{E}[\|G(\hat{w}, s) - x\|_1]) + \lambda_2(R(\hat{w}) + R(s) + R(r'))$	Simultaneously minimizes the rate, perception and distortion with hyperparameters λ_1 and λ_2
Cross-entropy/cosine embedding [76]	$L = L_{rec} + \lambda[1 - \cos(S, em)]$	Combination of traditional cross-entropy L_{rec} and cosine embedding loss
Normalized root mean squared error [78]	$\mathcal{L}_{NRMSE} = \frac{\sum_{t=1}^{T}(a_t - \hat{a}_t)^2}{\sum_{t=1}^{T} a_t^2}$	Seeks to minimize the difference between recovered audio data $\hat{a}_t$ and actual data a_t
Mean squared error [33]	$\mathcal{L}_{MSE} = \frac{1}{W}\sum_{w=1}^{W}(s_w - \hat{s}_w)^2$	Seeks to minimize the difference between recovered speech data $\hat{s}_w$ and actual data s_w

have been used to create artificial players which outperform human world champions in both games [86].

Recently, some have turned to using RL as a means to achieve semantic communication. In [87], a RL solution is proposed to carry out general semantic communication. It is argued that the objective functions of many of the ML-based approaches to semantic communication demonstrate a "semantic blindness," and are still biased toward bit-level accuracy. The proposed joint source-channel coding solution, termed SemanticRL-JSCC, is formulated using a Markov decision process framework. The reward function is based on any *general* semantic similarity function. This method differs from those DL methods discussed, as the semantics of the problem are defined by the chosen similarity function, and thus SemanticRL focuses on *how* to best communicate given some semantics, rather than learning *what* those semantics are. A distinct feature of SemanticRL as opposed to other ML approaches is that this similarity function is not necessarily differentiable. In the training of the model, a *self-critic* approach [88] is taken, resulting in a quicker and simpler solution. Experiments carried out over the European parliament dataset demonstrate superior performance with regard to common metrics, as well as a stable learning trajectory.

Another work combines a KG representation of semantic information with RL for semantic communication using text data [89]. First, a KG is extracted from a body of text, and this KG is treated as the semantic information of that text. Based on this KG representation, two metrics, namely semantic *accuracy* and *completeness* are derived. Combining these two metrics gives the overall metric of semantic similarity. Semantic communication is then formed as an optimization problem which seeks to maximize the semantic similarity of the text at the transmitter and receiver through resource allocation and information transmission, under a delay constraint. Using an attention-based RL framework to solve this optimization problem, it is shown that the proposed semantic communication solution outperforms a traditional RL scheme as well as typical wireless communication techniques.

Collaborative RL is a form of RL in which multiple agents are present in the system, and they collaborate to determine the optimum policy. In [90], collaborative deep RL (CDRL) is used to train a group of heterogeneous agents over a wireless cellular network. First, the algorithm selects the best subset of semantically relevant DRL agents for collaboration. This semantic relevance between two agents is based on their policies; if a target agent returns a large average reward under a source agent's policy, the target is said to be similar to the source. Here, the semantics are captured by the policies of each of the agents; similar meaning is implied by a similar policy. Once the similar subset of agents is obtained, the training loss and wireless bandwidth are jointly minimized to obtain the optimal policies for each agent. Simulations of the proposed technique show improved training performance compared to other CDRL methods and classic DRL. It is also shown that the proposed approach is able to use resources more efficiently, demonstrating better performance with fewer resource blocks than other approaches.

Looking to implement RL-based semantic communication for a specific application, [91] proposes a DRL framework for air-to-ground URLLC communcation using unmanned aerial vehicles (UAVs). Similar to [90], this work proposes

the use of a multi-agent DRL framework, coined graph attention exchange network (GAXNet), for semantic communication. Self-attention is used to determine the attention a UAV gives to other UAVs in the network, and based on this attention, training is performed in a centralized manner. Once the optimal policies have been obtained, the central unit distributes the model to the UAVs and actions are carried out in a decentralized manner. It is shown that the proposed GAXNet achieves more efficient training than the state-of-the-art centralized training and decentralized execution algorithm QMIX [92], and is better able to avoid collisions between UAVs.

C. SUMMARY

ML-based semantic communication is an approach that has seen a spike of interest with the recent boom in AI technology. In ML-based semantic communication, the semantics of the problem are not predefined as in classical and KG-based semantic communication, but rather they are learning through data-driven training. This learning is performed either through DL or RL. Many different DL models have been proposed to facilitate semantic communication, include the transformer, GAN, CNN, and others. One notable example is DeepSC, which was originally proposed as a text-based semantic communication system using a transformer-based network. Variants of DeepSC have been proposed in recent years which target different modes of communication. Though not as prominent, some RL methods have been proposed to learn semantics as well. One benefit of this approach is the loss function need not be differentiable. In ML-based semantic communication, meaning is characterized by the parameters of a model which are learned in a data-driven manner.

Many of the discussed works provide quantitative results demonstrating efficient semantic communication. The results of [73] show a 2-4 times reduction in bit rate, while experiments in [78] indicate large performance gains over traditional systems for around 1/3 the original bit rate. Reference [81] demonstrates a 20% reduction in transmitted symbols to achieve a similar performance as baseline systems. DeepSC and its variants also indicate potential for traffic reduction; DeepSC-S [33] can achieve improved speech performance for similar bit rates, and MU-DeepSC [34] achieves superior accuracy with a 70% reduction in transmitted symbols. Finally, L-DeepSC [32] reduces traffic in another way, by drastically reduce the parameters of a ML model which is to be broadcast to IoT devices. All in all, as in KG-based semantic communication, existing works in ML-based semantic communication indicate the potential for this approach to address the issues raised at the outset.

D. CHALLENGES AND OPPORTUNITIES

One of the challenges with ML-based approaches to semantic communication was pointed out in [87], which is the difficulty in working with semantic metrics. As most DL techniques use gradient-based methods for optimization, we must necessarily work with differentiable metrics, while many semantic similarity metrics are non-differentiable. Another critical challenge is the black-box nature of many DL models. A challenge within the field of DL as a whole, this quality of deep networks obscures our ability to analyze and evaluate *why* a model does or does not perform well. Finally, the improved performance that DL has enjoyed due to big data is approaching a limit, where futher gains can only be reached through massive training, which imposes a huge computational burden [30].

Opportunities can be found in the solutions to these challenges. First, there is certainly no ubiquitous metric of semantic similarity, and it is likely that many metrics will be application-specific. Development of such metrics is of critical importance to the advancement of semantic communication. As each approach to semantic communication we've seen thus far (classical, KG-based, ML-based) employs its own unique metrics, it is likely that a convergence of these ideas and combination of metrics could yield novel implementations and improved results. Furthermore, development of methods similar to [87] that relax the differentiability requirement of the semantic metric is a promising avenue for further study.

Regarding the lack of interpretability and ever-growing appetite for data associated with deep networks, one promising solution is the development of *model-based deep learning* techniques [30]. Model-based methods are those that derive some inference rule based on prior knowledge of the problem, while data-driven methods (including DL) rely solely on data to form the inference rule. Model-based methods are typically more interpretable and efficient than data-driven methods, while data-driven methods are more expressive and robust in new situations. Techniques utilizing the strengths of both methods to create a single model are referred to as model-based DL [30], and are another promising method of addressing some of the challenges facing ML today.

VI. SIGNIFICANCE-BASED SEMANTIC COMMUNICATION

One last view on semantic communications has been proposed only recently, and involves defining the semantics of information as the *significance* of this information [16]. Recall "The effectiveness problem" defined in Section I: "How effectively does the received meaning affect conduct in the desired way?" Defining the semantics of information as significance of information essentially addresses this problem, as significance is inherently determined by what one is trying to achieve with communication, or the *goal*. As a simple example, if the goal of communication is to control a robot performing remote surgery (a necessarily real-time application), information that was just obtained will be much more significant than information obtained 10 seconds ago. Based on this general idea, [16] calls for "a redesign of the entire process of information generation, transmission and usage in unison." In this section, we survey the few recent works that support this idea of significance-based semantic

TABLE 8. Summary of works in significance-based semantic communication.

Age of Information	Uysal *et al.* [16]	Age of information as a significance-based semantic metric
	Uysal *et al.* [93]	Practical evaluation of the age of information metric
	Beytur *et al.* [94]	Age of information performance of UDP and TCP protocols
	Ayan *et al.* [96]	Comparison of AoI and VoI impact on performance of cellular control system
Value of Information	Uysal *et al.* [16]	Value of information as a significance-based semantic metric
	Molin *et al.* [95]	Value-based information management for state estimation
	Ayan *et al.* [96]	Comparison of AoI and VoI impact on performance of cellular control system
Semantic Sampling	Uysal *et al.* [16]	Semantic sampling as part of a semantic communication architecture
	Kountouris & Pappas [17]	Semantic sampling as part of a semantic communication architecture
	Bacinoglu *et al.* [97]	Semantic sampling for tracking of a stochastic process
	Dommel *et al.* [98]	Semantic sampling for goal-oriented communication over a shared medium

communication. The works discussed in this section are summarized in Table 8.

In [16], some examples of measures which relate to the significance of information are given. The first of these is information *freshness*, which is determined by the time taken since the information was generated to when it was received. Age of Information (AoI) is a measure which captures this idea of freshness, and has been well-studied over the past decade or so. We will discuss AoI further in the following subsection as a prime example of a significance-based semantic measure; [16] presents results indicating energy savings of different age-aware protocols ranging from 10-64%. Another example of a semantic measure is *relevance*. Consider a process that is being sampled; consecutive samples that capture little change in the process are typically of less interest than those for which sudden changes occur. We could say that the latter samples are more relevant than the former. As an extension of relevance of information, a more powerful example is given as the *value* of information, which is defined as the difference between the benefit of a sample and the cost of its transmission. It is argued in [16] that developing metrics that capture these ideas is critical for achieving semantic communication. Those mentioned here are summarized in Table 9.

Reference [16] also presents a vision for an end-to-end semantic communication architecture. This architecture takes into considerations the elements of freshness, relevance, and value to optimize the entire system. A simplified flowchart of the proposed architecture is provided in Figure 13. Specifically, semantic sampling is implemented to relax the assumption that data arrives in an uncontrolled manner, i.e., only significant information is generated in the first place. Semantic channel encoding, multiuser scheduling, channel access and flow control are proposed to increase the efficiency and effectiveness of each of these processes. It is stated that the realization of this architecture will involve an entirely new paradigm shift that is incompatible with previous designs of communication systems. A few specific examples are envisioned, which include semantic communication for networked control systems, smart cities, and mMTC/IoT systems.

Reference [17] is a seemingly independent work from that of [16] which emerged around the same time and shares the idea of significance-based semantic communication. Reference [17] defines semantics of information as "the significance and usefulness of messages." A similar argument to that of [16] is given, stating that simple generation and communication of data often leads to reception of stale or irrelevant information at the receiver and wasted resources. This again brings about the need for a *goal-oriented* communication system, one that addresses "The effectiveness problem."

A new concept presented in [17] is the idea of defining semantics at different *scales*. At the *microscopic* scale, specific pieces of information from the source may be of different significance, e.g., that a safety risk is present or not. The *mesoscopic* scale is the intermediate level, which takes into consideration link-level semantics. This includes both innate (objective) measures such as freshness (AoI) and precision, and contextual (subjective) measures such as timeliness and completeness. Finally, the *macroscopic* scale takes system-level semantics into consideration, specifically looking at end-to-end distortions and delays that affect the end goal.

An end-to-end semantic architecture is proposed, which is strikingly similar to that envisioned in [16], including semantic sampling and semantics-aware signal processing blocks. A specific example of an end-to-end communication system is given, involving a remote actuation application. The source monitors the actual state of a robotic arm, while the receiver aims to construct and maintain a digital twin of this robotic arm. Communicating over a wireless erasure channel, it is shown that an end-to-end semantics approach performs much better than other, more semantically-unaware approaches in terms of real-time reconstruction error and cost of actuation error.

TABLE 9. Summary of significance-based semantic measures [16].

Measure	Expression	Description
Age of Information	$\Delta(t) = t - u(t)$	Characterizes the *freshness* of information generated at time $u(t)$
Value of Information	$v(x) = b(x) - c(x)$	Difference between the benefit of a sample and cost of its transmission
Relevance of Information	$r(x) = f(x_n, x_{n-1})$	Quantifies the amount of change in a process since the previous sample

A. AGE OF INFORMATION AND VALUE OF INFORMATION

Two proposed measures that capture the significance of information are AoI and VoI. In this subsection, we will discuss some of the work that has been done with regards to these two measures to illustrate their use as semantic measures.

Reference [93] provides a compilation of some recent works examing AoI in practical scenarios and looks at issues such as synchronization, transport layer protocols, congestion, and the use of ML. First, the status *age* of an information flow (characterized as a flow of data packets) is defined as the difference between the current time and the generation time of the most recently received data packet, which is illustrated in Figure 12. Other derivative metrics are defined as well, such as *average* AoI and *peak* AoI. Depending on the application at hand, one may desire to minimize either the average or peak AoI to best maximize information freshness.

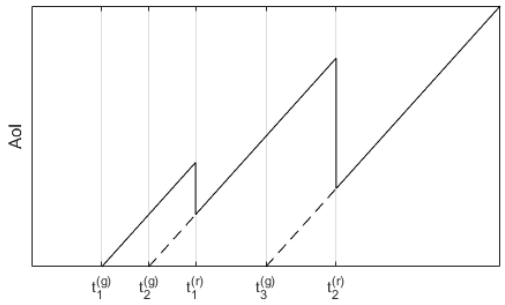

FIGURE 12. Example of AoI, where $t_n^{(g)}$ and $t_n^{(r)}$ is the generation and reception time of the *n*th packet, respectfully.

Next, [93] discusses the measurement of AoI in practical systems. It is noted that first and foremost, accurate timing information is required for age measurement. Second, synchronization is required at the transmitter and receiver. Once these have been established, AoI can be computed at the transmitter, receiver, or centrally and used for network optimization. It is also shown how timing imperfections, such as clock bias, can affect AoI measurement.

Reference [93] then looks at a specific work which examines the AoI performance of some modern transport layer protocols. In [94], the AoI performance of User Datagram Protocol (UDP) and Transmission Control Protocol (TCP) are examined for different testbed setups. It is shown that UDP is able to maintain a lower average AoI at higher data rates than TCP for a multi-hop network testbed. Both protocols tend to perform relatively well up to a certain rate, above which the system becomes "panicked" and AoI performance becomes poor. In contrast, evaluated with respect to an IoT tesetbed, the opposite holds true, and TCP is shown to achieve a slightly better AoI performance. Overall, [93] provides a useful summary of how AoI can be practically integrated to evaluate communication systems.

As an example application of VoI, [95] proposes a value-based method of information management of a networked system for state estimation. Essentially, this system will allocate a time slot to the estimator with the highest-priority information, where priority is determined by the VoI. The VoI of each estimator is computed as a function of the expected overall weighted squared error, given the current data at that local estimator. Therefore, by choosing the data which minimizes this error, the system is essentially choosing the information with the highest value to the task at hand. To illustrate the performance of the proposed system, an automated driving scenario with multiple vehicles is simulated. It is shown that the VoI-based scheme is able to avoid collisions with very high probability, while a simple time-triggered scheduling approach resulted in a collision in 19.7% of the experiments.

Another work has directly compared the impact of AoI and VoI on the performance of a cellular networked control system [96]. Here, VoI is defined as quantifying the amount of reduction in uncertainty of a stochastic process at the recipient. It is interesting to note the similarity between this definition and that of Shannon's definition of entropy [5], which also quantifies uncertainty reduction. Here, VoI is concerned with the content of a new update, while AoI focuses only on the timeliness of this update. The AoI is defined in the usual fashion, while a VoI metric is proposed for both uplink and downlink transmissions, and in both cases is a function of the expected squared-error. Simulations demonstrate that a system implementing a VoI-based scheduler is able to achieve a lower absolute error as opposed to a system with an AoI-based schedule for the cellular networked control system.

B. SEMANTIC SAMPLING

Another important aspect of a significance-based semantic communication system, as proposed by both [16] and [17] is the idea of *semantic sampling*. Basically, the aim of semantic sampling is to generate information at the source in a "smart" way, such that only necessary information is generated and transmitted over the system. As a general example of this, [97] considers the problem of tracking an unstable

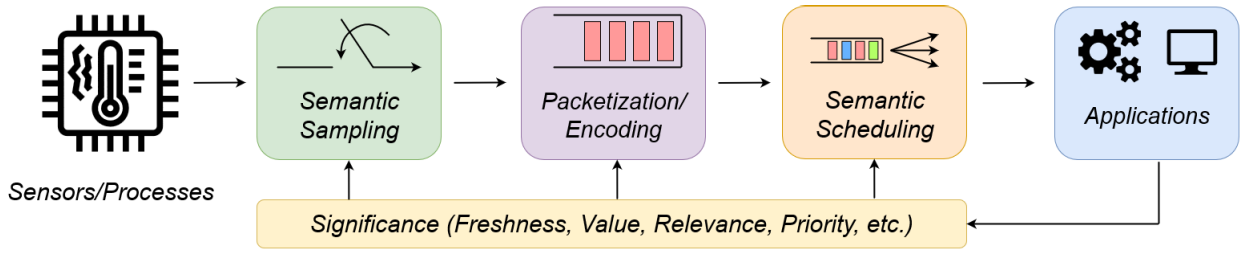

FIGURE 13. Overview of the significance-based semantic architecture proposed in [16].

stochastic process by using causal information of another stochastic process. Essentially, by using some information related to the process of interest, we can determine when to take "significant" samples that allow for accurate tracking, thereby implementing semantic sampling. This work can be seen as contributing to a theory of semantic sampling as discussed in [16]. In [97], necessary conditions are provided for tracking integer-valued sources using causal information. These results are expressed in terms of the Rényi entropy and information density; essentially, the information density between the two processes must be greater than a threshold that is set by the Rényi entropy of the process being tracked. Furthermore, [97] also provides sufficient conditions for tracking integer-valued sources using causal information. The first of these is based on MAP estimator of the source information based on the causal information, and the second is based on a different estimator which considers a notion of distance. With regards to semantic sampling, the results of [97] imply that one could perhaps only sample and transmit the causal information over the channel instead of the source information itself. If the causal information results in fewer transmitted symbols, this would theoretically increase the efficiency of our system while preserving reconstruction fidelity at the receiver.

Looking at a more practical implementation of semantic sampling, [98] addresses the problem of semantics-aware active sampling and transmission over a shared communication medium. The goal of the system is to use joint sampling and transmission to compute the probability of a quantity of interest at the receiver. In this work, semantics-aware communication refers to a system in which the receiver aims to recover the aggregated information of interest, rather than the individual messages. For example, perhaps the average measurement from a number of sensors is of interest; in this case, meaning is captured by the aggregated value rather than the individual data. An active sampling scheme is adopted, such that each device takes samples according to a Bernoulli distribution, where the parameter of this distribution can vary between devices. To obtain the empirical probability of the quantity of interest, each device transmits over a time slot, and the receiver averages over the time slots. The estimation technique is shown to perform well with respect to both mean squared-error and Kullback-Leibler Divergence metrics.

C. SUMMARY

One final approach to semantic communication is significance-based communication, which addresses both the semantic and the effectiveness problems of communication. This recent approach makes use of metrics which quantify the significance of information, such as freshness, value, relevance, and others. Two prime examples of metrics corresponding to significance-based semantic communication are AoI and VoI. AoI is a popular metric which has been well-studied compared to VoI, and some work has been done comparing the two measures. Furthermore, a key idea of this approach is semantic sampling, which has received some attention as well. Overall, significance-based semantic communication looks to solve the semantic problem by first solving the effectiveness problem, and assigning meaning to information based on what impact that information will have at the receiver.

As the most recent of the discussed methods, quantitative results demonstrating the potential of significance-based semantic communication for data traffic reduction are few. In [16], results are presented demonstrating energy savings, resulting in decreased transmission, of 10-64% for different age-aware protocols. While there are not many quantitative results in this area, the potential is clear. Significance-based communications center around the reduction of insignificant information, and thereby inherently work to reduce traffic in a communication system.

D. CHALLENGES AND OPPORTUNITIES

One clear challenge faced by significance-based semantic communication, similar to other methods of semantic communication, is the development of applicable metrics. AoI and VoI are two examples which have been the focus of prior work, however we anticipate that other useful metrics will be proposed as this approach to semantic communication is being developed in the literature. The quest for appropriate metrics presents a rich opportunity for future research. Clearly these measures are highly application-specific, and thus the utility of one measure may vary greatly from one scenario to another.

Another challenge relates to the general progression to such a semantic communication system. As stated in [16], this approach entails a radical departure from the ways in which current communication systems operate. However, to be a

viable path forward for modern communication systems, a certain level of backward-compatibility must be present, such that the vast existing wireless infrastructure need not be replaced from scratch. Methods which involve some degree of compatibility with legacy systems will be important for the progression to semantic communication.

Many opportunities lie in the development of significance-based semantic communication systems. With the advancement of IoT and cyber-physical systems, wireless communication is increasingly being used for highly specific goal-oriented tasks. Designing systems which communicate as efficiently as possible under the constraints imposed by the specific task at hand will be important for optimizing the efficiency of wireless technologies, as well as improving performance of these technologies.

VII. A DIFFERENT APPROACH: CONTEXT-BASED SEMANTIC COMMUNICATION

Throughout this survey, we have presented a review of the history and state of the art of semantic communication by examining the different approaches toward engineering this higher level of communication. Each of these approaches differ in how they treat the semantics of the problem. Classical approaches attempt to quantify semantic information in probabilistic terms, much the same way as traditional information theory. KG-based semantic communication uses KGs to represent knowledge of the semantic source and receiver, from which semantic methods can be derived. In ML-based semantic communication, data is used to learn the latent ''semantic'' relationships and optimize communication based on these relationships. Significance-based semantic communication essentially combines ''The semantic problem'' with ''The effectiveness problem'' and emphasizes efficient, goal-oriented communication.

In this section, we present a novel approach to the semantic communication problem. We first argue that *context* is at the heart of all semantic communications. While context is certainly an implicit consideration in each of the aforementioned approaches, we believe than an explicit and deliberate focus on the context of communication will lead to novel and valuable semantic communication systems. We then present our view of how to define context, and our vision of a systematic design procedure which frames semantic communication as a context-dependent, goal-oriented optimization problem.

A. THE IMPORTANCE OF CONTEXT

To motivate the utility of a context-based approach, recall again the example from Section I, where a speaker wants to communicate how to compute the area of a circle to a listener. Semantic communication in scenario 1 (listener vaguely familiar with geometrical concepts) can be much more *syntactically* efficient than the same semantic communication in scenario 2 (listener is a small child). As was illustrated in Section I, the key observation is that the listener in either scenario starts with a different prior knowledge base.

However, it is important to note that this is not the *only* characteristic of the scenario which will affect communication. What if scenario 1 takes place in a one-on-one office meeting, while scenario 2 takes place in a crowded classroom of restless children of similar age? Certainly the efficiency-of-communication gap between the two scenarios will widen. Furthermore, say that the speaker has a one-hour one-on-one office meeting with the listener in scenario 1, but has four 15-minute sessions in the crowded classroom with the young listener of scenario 2 held on different days. This again will impact how to most efficiently communicate in each scenario; it is likely that some review will be needed in each of the disjoint sessions of scenario 2.

Take any situation in which communication occurs, and a similar analysis can be done to determine the factors that impact the way in which communication is carried out. Based on this observation, the characterization of these factors is clearly an important step in designing an efficient semantic communication system. While context is inherent in any approach to semantic communication, the previously discussed methods only implicitly consider this key factor. Based on the above example, we argue that an explicit focus on context is needed for optimal semantic communication.

B. DEFINING CONTEXT

We broadly refer to any factors that impact *how* one efficiently communicates as being part of the *context* of the problem. In the example, one factor was the parties involved in the communication process. Another was the setting, i.e., a quiet office vs. a noisy classroom. The third factor involved temporal aspects of the situation, i.e., a long, uninterrupted session vs. short sessions spanning multiple days. We observe that these factors each correspond to difference pieces of the overall context (people, place, and time). We therefore postulate that context can be completely described by considering what are sometimes referred to as the ''five Ws'':

- *Who*: Agents involved in the communication process
  - Includes source(s), receiver(s), and/or other agents that are involved indirectly
- *What*: The mode of communication
  - Could be text, speech, etc.
- *Where*: Qualities of the environment in which communication occurs
  - Specifies channel characteristics
- *When*: Temporal aspects of the problem/environment
  - Considers static vs. dynamic agents, mode, channel, and goal
- *Why*: The purpose of communication
  - Defines what is to be achieved

These aspects of context can in turn be incorporated into the mathematical model of the communication problem at hand. For example, the *What* aspect dictates the space of signals or symbols available to the source for communication, the *Where* aspect defines which channel model is to be used, and the *Why* aspect may be some function which specifies quality of service or other desired outcomes. By explicitly

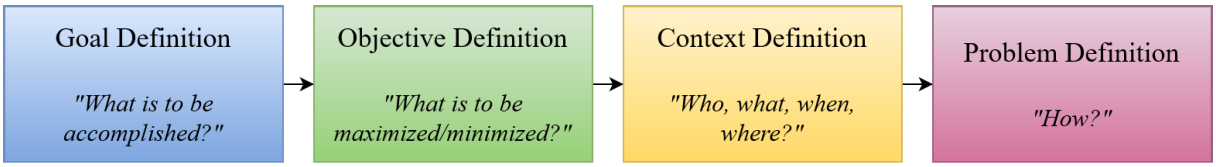

FIGURE 14. Proposed design flow for context-based semantic communication systems.

considering *Why* as part of the context, this is similar to significance-based communication in that it also addresses "The effectiveness problem." Another term for this is *goal-oriented communication*. Once a mathematical description of each aspect of the context is available, they can be incorporated into an optimization framework to achieve efficient semantic communication. This contextual knowledge may be available *a priori*, or it may need to be learned online using modern data-driven techniques. In either case, just like the speaker in the example, an optimal communication strategy can be devised by explicitly taking these different aspects into consideration.

C. CONTEXT-BASED DESIGN

Using context as it is defined above, we propose a general design procedure for semantic communication systems. Note that we use the term "context-based" to imply that some optimization is being carried out based on explicit consideration of the context. The procedure, illustrated in Figure 14, involves the systematic construction of an optimization problem and consists of the following steps:

1) *Goal definition*: First, the *why* aspect of the context is determined. As is true in any engineering setting, we must first know what problem we are solving; what is it we are trying to accomplish? This will be a constraint in the overall optimization problem.
2) *Objective definition*: This step will determine exactly what is to be optimized, e.g., energy efficiency, spectral efficiency, etc. As the name implies, this will determine the objective function to be optimized.
3) *Context definition*: Now, define the remaining aspects of the context, namely *who*, *what*, *where*, and *when*. These will also manifest as constraints in the optimization problem.
4) *Problem definition*: Define a set of communication *strategies* to be optimized over. Use this set, along with the objective function and constraints derived from steps 1-3, to define the optimization problem.

As a general example, suppose we first obtain a goal represented by a constraint C_G. An objective function f that we wish to minimize is identified, followed by N context-based constraints $C_1, C_2, \ldots, C_N$. Finally, a set of possible communication strategies is determined and denoted by $\mathcal{S}$. Then, context-based semantic communication is performed by selecting a strategy as a solution to the optimization problem defined by

$$\min_{S\in\mathcal{S}} . \quad f(S) \tag{29}$$

$$\text{s.t.} \quad C_G, C_1, \ldots, C_N \tag{30}$$

Depending on the characteristics of the resulting problem, how to solve it becomes a challenge in itself. If the resulting problem is convex, then well-known methods of solving are readily available [99]. In the likely case that the problem is non-convex, the problem becomes harder to solve, in which methods such as convex relaxation or ML may be needed to make the problem tractable.

As a concrete example, consider smart agriculture, which is expected to play a prominent role in the 6G network [100]. Furthermore, suppose we desire to accurately monitor soil moisture in the field using context-based semantic communication between a set of J sensors and a single fusion node (FN). Following the framework above, the goal is to produce an accurate picture of the soil moisture throughout the field; mathematically, this can be expressed by a metric called *confident information coverage* [101]:

$$\Phi(x) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(z^t(x) - \hat{z}^t(x))^2} \tag{31}$$

where $z^t(x)$ and $\hat{z}^t(x)$ are the actual and estimated soil moisture values at point x and time t, respectively, and T is the time period over which estimation takes place. We say that the field is "completely confident information covered" if $\Phi(x) < \epsilon$ for all x in the field $\mathcal{X}$. We take this to be our goal, and correspondingly the first constraint of the optimization problem.

Next, we must define the objective. Suppose that we are interested in maximizing the lifetime of the sensor network, and thus minimizing the power of the sensors, as is typical in an IoT application. Suppose that associated with sensor j is a set of sensing powers $\mathcal{P}^{(j)}_{\text{sense}}$ and transmit powers $\mathcal{P}^{(j)}_{TX}$. Intuitively, we assume that greater sensing power will produce more accurate sensing, and greater transmit power will produce higher quality communication. Then the total power of the sensors is given by

$$P = \sum_{j=1}^{J}(P^{(j)}_{\text{sense}} + P^{(j)}_{TX}), \tag{32}$$

where $P^{(j)}_{\text{sense}} \in \mathcal{P}^{(j)}_{\text{sense}}$ and $P^{(j)}_{TX} \in \mathcal{P}^{(j)}_{TX}$ for all $j \in \{1, 2, \ldots, J\}$. We can now also define a set of communication

strategies as $\mathcal{S} = \times_{j\in J}(\mathcal{P}_S^{(j)} \times \mathcal{P}_T^{(j)})$, and the objective becomes

$$\min_{S\in\mathcal{S}} \quad P = \sum_{j=1}^{J}(P_{\text{sense}}^{(j)} + P_{TX}^{(j)}). \tag{33}$$

To define the context, we must consider the "four Ws" listed above. We will use a state-space representation to define the context. To address the *Who* question, let $\Omega_s = \{\omega_1, \omega_2, \ldots, \omega_J\}$ and $\Omega_r = \{\omega_r\}$ represent the states of the sensors and the FN, respectively. For example, ω_i could indicate whether sensor i is online or offline, and ω_r might indicate whether the FN is receiving or computing. *What* refers to the symbols used to communicate; e.g., under some strategies a sensor may use more precise quantization than others, resulting in longer symbols. *Where* will encompass the channel effects between each sensor and FN, and can be represented by $\Omega_c = \{\omega_1, \omega_2, \ldots, \omega_J\}$. Taking the cross product of the individual state-spaces gives the overall state-space $\Omega = \{\Omega_s \times \Omega_r \times \Omega_c\}$. Finally, the *When* question is addressed by temporal changes in the state space. Assuming that we observe the state at discrete time instances, this can be expressed by representing the state-space as a function of time $\Omega[n]$, $n = 1, 2, \ldots, \infty$.

Putting all of this together as the final step in the process, we arrive at the context-based optimization problem

$$\min_{S\in\mathcal{S}}. \quad P = \sum_{j=1}^{J}(P_S^{(j)} + P_T^{(j)}) \tag{34}$$

$$\text{s.t.} \quad \Phi(x) < \epsilon, \ \ \forall\, x \in \mathcal{X} \tag{35}$$

$$\Omega = \Omega[n]. \tag{36}$$

By framing communication in this optimization framework, the system is acting as the teacher from our initial example, namely by communicating with the underlying goal of efficiency, which is achieved by considering the context in which communication is taking place. As mentioned above, this formulation only presents us with a problem for which solving is another matter. As our goal here is to introduce the framework itself, we leave further study of this second stage of the problem for future work.

Semantic communication is regarded as a promising solution for improving the efficiency of communication systems. More so than the previously discussed techniques, the proposed context-based method is formulated with this specific aim in mind. By considering this aim at the outset, and taking into consideration the context of the communication problem, we believe that the resulting semantic communication systems will have the potential to advance the state of the art once more.

VIII. CONCLUSION

The aim of this survey is to provide a comprehensive and clear picture of the current state of the emerging field of semantic communications. The push toward semantic communication systems is motivated by the explosion of global data traffic demand in recent years, and the intuitive benefits that can be achieved through efficient communications. Defining semantics is a non-trivial problem, and some approaches have emerged in the literature.

Classical semantic information-based approaches attempt to extend the ideas of information theory to capture the semantics of information, and are based on ideas of logical probabilities and truthlikeness. Two prominent theories are TWSI and TSSI, and truthlikeness-based approaches extend these theories. The key idea of this approach is to follow the path of classical Information Theory by first quantifying semantic information, and developing results based on this quantification.

KG-based semantic communication focuses on the aspect of a knowledge base at a semantic source and receiver, using a KG to model such knowledge bases. This approach follows from the extensive work revolving around the semantic web. By representing knowledge in a graph structure, semantic similarity measures can be devised and forms of reasoning can be performed. This form of structure and working with knowledge is the key driver of this approach.

ML-based semantic communications uses modern learning techniques to carry out semantic communication in a data-driven manner. This includes DL methods, which learn semantics through neural network structures, and RL techniques, which learn semantics with an action-reward framework. Unlike the classic and KG-based approaches, this approach assumes no formal structure of the semantics to be learned, and puts the burden of learning these semantics on the model itself. Consequently, meaning is captured by the tuned model parameters that are found as a result of data-driven training.

Finally, significance-based semantic communications take a goal-oriented approach and look to communicate in a way that best achieves the goal. Significance is quantified by metrics pertaining to different qualities of information, such as freshness and value. Semantic sampling is also a critical point of this approach, which seeks to generate only information that is pertinent to the task at hand. By addressing the effectiveness problem of communication, this approach circumvents the semantic problem and inherently assumes that the semantics will be addressed within the effectiveness solution.

Regarding the problem of data traffic reduction, various works in KG-based and ML-based semantic communication have demonstrated quantitative results illustrating the potential to address this growing issue for diverse applications and use-cases, from simple text-based speech to network operations recommendation systems. For the classical and significance-based approaches, these results are fewer. One conclusion that can be drawn upon examination of these various results, is the need for standard evaluation procedures across the different approaches. If efficient communication is to be one of the main goals of semantic communication, standard metrics capturing this idea should be chosen as the field develops to facilitate the comparison between competing methods.

For each approach, we provide some challenges and opportunities that could inspire future work in the field. For the classical approach, a clear opportunity lies in the further development of a theory of semantic information. The KG-based approach includes many existing methods which can be extended; particularly, scalability and learning methods are two areas which require further improvements. A major drawback of the ML-based approach is a lack of interpretability, and model-based DL and neurosymbolic AI are two potential solutions to this issue. The significance-based approach is the most recent of the four, and thus provides many exciting opportunities for future work, particularly in the development and study of significance-oriented metrics. The most important challenge to the success of semantic communication is the ability to define and work with ''meaning.'' We believe that future semantic communication systems will leverage techniques across the different approaches to optimize these systems, and thus we view future work corresponding to each approach as important to the overall field.

Furthermore, we advocate for a fifth approach to engineering semantic communication, namely context-based semantic communication. This approach places an emphasis on the context of the communication problem, which in turn impacts the strategy that leads to efficient communication. Based on the observation that humans naturally optimize communication as a result of the context, our approach involves the formulation of the strategy selection as an optimization problem, which can be solved using traditional or modern techniques. We demonstrate the details of this approach with a smart agriculture example based on a soil moisture monitoring application.

Realizing this higher level of communication is an exciting problem that presents a plethora of challenges and opportunities for future work. In the age of ever-expanding AI and ML, it is only natural that we apply this intelligence to communication systems to reap the benefits therein. Our hope is that this survey will prove to be a useful guide to anyone interested in the engineering of semantic communication.

REFERENCES

[1] Ericsson. (2021). *Ericsson Mobility Report*. [Online]. Available: https://www.ericsson.com/4ad7e9/assets/local/reports-papers/mobility-report/documents/2021/ericsson-mobility-report-november-2021.pdf

[2] Organisation for Economic Co-Operation and Development. (2021). *Teleworking in the COVID-19 Pandemic: Trends and Prospects*. [Online]. Available: https://www.oecd.org/coronavirus/policy-responses/teleworking-in-the-covid-19-pandemic-trends-and-prospects-72a416b6/

[3] 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.

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

[5] C. E. Shannon, ''A mathematical theory of communication,'' *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[6] Federal Communications Commission Office of Engineering and Technology Policy and Rules Division. (2021). *FCC Online Table of Frequency Allocations*. [Online]. Available: https://www.fcc.gov/file/21474/download

[7] A. Mchangama, J. Ayadi, V. P. G. Jimenez, and A. Consoli, ''mmWave massive MIMO small cells for 5G and beyond mobile networks: An overview,'' in *Proc. 12th Int. Symp. Commun. Syst., Netw. Digit. Signal Process. (CSNDSP)*, Jul. 2020, pp. 1–6.

[8] E. C. Strinati, D. Belot, A. Falempin, and J.-B. Dore, ''Toward 6G: From new hardware design to wireless semantic and goal-oriented communication paradigms,'' in *Proc. IEEE 47th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2021, pp. 275–282.

[9] R. Carnap and Y. Bar-Hillel, ''An outline of a theory of semantic information,'' *J. Symbolic Log.*, vol. 19, no. 3, pp. 230–232, 1954.

[10] *Engineering, N*. Accessed: Sep. 1, 2022. [Online]. Available: https://www-oed-com.er.lib.k-state.edu/view/Entry/62227?result=2&rskey=ARkCIq&

[11] M. A. Ouksel and C. F. Naiman, ''Toward the design of a semantic communication protocol in heterogeneous database systems,'' in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 2, Oct. 1992, pp. 1271–1276.

[12] T. Berners-Lee, J. Hendler, and O. Lassila, ''The semantic web,'' *Sci. Amer.*, vol. 285, no. 5, pp. 34–43, 2001.

[13] V. Rodoplu and S. S. Vadvalkar, ''Challenges and directions for semantic communication,'' in *Proc. Int. Conf. Semantic Comput.*, Sep. 2007, pp. 290–294.

[14] B. Juba, ''Compatibility among diversity foundations, lessons, and directions of semantic communication,'' in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, Mar. 2013, pp. 458–463.

[15] B. Juang, ''Quantification and transmission of information and intelligence—History and outlook [DSP history],'' *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 90–101, Jul. 2011.

[16] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, T. Soleymani, B. Soret, and K. H. Johansson, ''Semantic communications in networked systems: A data significance perspective,'' *IEEE Netw.* vol. 36, no. 4, pp. 233–240, Jul. 2021.

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

[18] W. Tong and G. Y. Li, ''Nine challenges in artificial intelligence and wireless communications for 6G,'' *IEEE Wireless Commun.*, vol. 29, no. 4, pp. 140–145, Aug. 2022.

[19] E. C. Strinati and S. Barbarossa, ''6G networks: Beyond Shannon towards semantic and goal-oriented communications,'' *Comput. Netw.*, vol. 190, May 2021, Art. no. 107930.

[20] 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.

[21] 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.

[22] B. Juba, *Universal Semantic Communication*. Berlin, Germany: Springer, 2011. [Online]. Available: https://books.google.com/books?id=L9Yay8hlqikC

[23] J. Dewey, *Experience and Nature* (Lectures Upon the Paul Carus Foundation). Tempe, Arizona: Norton, 1929. [Online]. Available: https://books.google.com/books?id=L_DWAAAAMAAJ

[24] L. Wittgenstein, *Philosophical Investigations*. Oxford, U.K.: Basil Blackwell, 1953.

[25] J. Delgado-Frias and W. Moore, ''A semantic network architecture for artificial intelligence processing,'' in *Proc. IEEE Int. Workshop Tools Artif. Intell.*, Jan. 1989, pp. 162–167.

[26] R. Rada, H. Mili, E. Bicknell, and M. Blettner, ''Development and application of a metric on semantic nets,'' *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 17–30, Jan./Feb. 1989.

[27] J. S. Mertoguno and W. Lin, ''Distributed knowledge-base: Adaptive multi-agents approach,'' in *Proc. IEEE Int. Joint Symposia Intell. Syst.*, Nov. 1996, pp. 76–83.

[28] 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. 2021.

[29] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, ''A comprehensive survey on graph neural networks,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Mar. 2020.

[30] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, ''Model-based deep learning,'' 2020, *arXiv:2012.08405*.

[31] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
[32] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, Jan. 2021.
[33] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
[34] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 553–557, Mar. 2022.
[35] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.
[36] L. Floridi, "Outline of a theory of strongly semantic information," *Minds Mach.*, vol. 14, no. 2, pp. 197–221, May 2004.
[37] S. D'Alfonso, "On quantifying semantic information," *Information*, vol. 2, no. 1, pp. 61–101, Jan. 2011.
[38] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in *Proc. IEEE Netw. Sci. Workshop*, Jun. 2011, pp. 110–117.
[39] P. Basu, J. Bao, M. Dean, and J. Hendler, "Preserving quality of information by using semantic relationships," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, Mar. 2012, pp. 58–63.
[40] A. A. Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd wave," 2020, *arXiv:2012.05876*.
[41] B. Swartout, R. Patil, K. Knight, and T. Russ, "Toward distributed use of large-scale ontologies," in *Proc. 10th Banff Knowl. Acquisition Workshop*, 1997, pp. 138–148.
[42] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneijder, and L. A. Stein, "OWL web ontology language reference," in *Proc. World Wide Web Consortium (W3C)*, 2004, pp. 1–15. [Online]. Available: http://www.w3.org/TR/owl-ref/
[43] P. Resnik, "Using information content to evaluate semantic similarity in a taxanomy," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, pp. 448–453.
[44] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," 1997, *arXiv:cmp-lg/9709008*.
[45] D. Sathya and K. R. Uthayan, "Proposal for semantic metric to assess the quality of ontologies," in *Proc. Int. Conf. Signal Process., Commun., Comput. Netw. Technol.*, Jul. 2011, pp. 754–756.
[46] A. Sheth, C. Henson, and S. S. Sahoo, "Semantic sensor web," *IEEE Internet Comput.*, vol. 12, no. 4, pp. 78–83, Jul./Aug. 2008.
[47] World Wide Web Consortium. (2015). *RDFA Core 1.1*. [Online]. Available: https://www.w3.org/TR/rdfa-core/
[48] J. Li, "Automatic semantic analysis framework of Dickinson's portfolio based on character recognition and artificial intelligence," in *Proc. 4th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Nov. 2020, pp. 1660–1664.
[49] A. Gyrard, C. Bonnet, and K. Boudaoud, "Enrich machine-to-machine data with semantic web technologies for cross-domain applications," in *Proc. IEEE World Forum Internet Things (WF-IoT)*, Mar. 2014, pp. 559–564.
[50] S. Chun, S. Seo, B. Oh, and K.-H. Lee, "Semantic description, discovery and integration for the Internet of Things," in *Proc. IEEE 9th Int. Conf. Semantic Comput.*, Feb. 2015, pp. 272–275.
[51] L. B. Bhajantri and R. Pundalik, "Data processing in semantic sensor web: A survey," in *Proc. 3rd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Dec. 2017, pp. 166–170.
[52] D. Schachinger and W. Kastner, "Semantic interface for machine-to-machine communication in building automation," in *Proc. IEEE 13th Int. Workshop Factory Commun. Syst. (WFCS)*, May 2017, pp. 1–9.
[53] E. Lakka, N. E. Petroulakis, G. Hatzivasilis, O. Soultatos, M. Michalodimitrakis, U. Rak, K. Waledzik, D. Anicic, and V. Kulkarni, "End-to-end semantic interoperability mechanisms for IoT," in *Proc. IEEE 24th Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Sep. 2019, pp. 1–6.
[54] M. Jeong, B. Kim, and G. G. Lee, "Semantic-oriented error correction for spoken query processing," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Nov. 2003, pp. 156–161.
[55] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," 2019, *arXiv:1905.07129*.
[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423
[57] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems*, vol. 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2013.
[58] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua, "Explainable reasoning over knowledge graphs for recommendation," 2018, *arXiv:1811.04540*.
[59] E. Aumayr, M. Wang, and A.-M. Bosneag, "Probabilistic knowledge-graph based workflow recommender for network management automation," in *Proc. IEEE 20th Int. Symp. 'World Wireless, Mobile Multimedia Netw.' (WoWMoM)*, Jun. 2019, pp. 1–7.
[60] H. He, A. Balakrishnan, M. Eric, and P. Liang, "Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings," 2017, *arXiv:1704.07130*.
[61] B. Guler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 4, pp. 787–802, Dec. 2018.
[62] Y. Wei, J. Luo, and H. Xie, "KGRL: An OWL2 RL reasoning system for large scale knowledge graph," in *Proc. 12th Int. Conf. Semantics, Knowl. Grids (SKG)*, Aug. 2016, pp. 83–89.
[63] D. Zheng, X. Song, C. Ma, Z. Tan, Z. Ye, J. Dong, H. Xiong, Z. Zhang, and G. Karypis, "DGL-KE: Training knowledge graph embeddings at scale," 2020, *arXiv:2004.08532*.
[64] H. Zhu, X. Wang, Y. Jiang, H. Fan, B. Du, and Q. Liu, "FTRLIM: Distributed instance matching framework for large-scale knowledge graph fusion," *Entropy*, vol 23, no. 5, p. 602, May 2021.
[65] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2020, *arXiv:2003.01200*.
[66] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," 2021, *arXiv:2111.07624*.
[67] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100379. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013721000198
[68] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC, USA: Spartan Books, 1962. [Online]. Available: http://catalog.hathitrust.org/Record/000203591
[69] Z. Li, W. Yang, S. Peng, and F. Liu, "A survey of convolutional neural networks: Analysis, applications, and prospects," 2020, *arXiv:2004.02806*.
[70] Y. Yu, X. Si, C. Hu, and Z. Jianxun, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, 2019.
[71] Y. Lu, Y. Shi, G. Jia, and J. Yang, "A new method for semantic consistency verification of aviation radiotelephony communication based on LSTM-RNN," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Oct. 2016, pp. 422–426.
[72] Y. Hua and J. Du, "Deep semantic correlation with adversarial learning for cross-modal retrieval," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 256–259.
[73] D. Huang, X. Tao, F. Gao, and J. Lu, "Deep learning-based image semantic coding for semantic communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.
[74] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
[75] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar. 2009.
[76] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "SEED: Semantics enhanced encoder–decoder framework for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13525–13534.

[77] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.

[78] H. Tong, Z. Yang, S. Wang, Y. Hu, W. Saad, and C. Yin, "Federated learning based audio semantic communication over wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[79] Q. Zhou, R. Li, Z. Zhao, C. Peng, and H. Zhang, "Semantic communication with adaptive universal transformer," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 453–457, Mar. 2022.

[80] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. Mach. Transl. Summit X, Papers*, Phuket, Thailand, 2005, pp. 79–86. [Online]. Available: https://aclanthology.org/2005.mtsummit-papers.11

[81] M. Sana and E. C. Strinati, "Learning semantics: An opportunity for effective 6G communications," in *Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2022, pp. 631–636.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[83] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[84] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1988–1997.

[85] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," 2017, *arXiv:1708.05866*.

[86] M. L. Littman, I. Ajunwa, G. Gerger, C. Boutilier, M. Currie, F. Doshi-Velez, G. Hadfield, M. C. Horowitz, C. Isbell, H. Kitano, K. Levy, T. Lyons, M. Mitchell, J. Shah, S. Sloman, S. Vallor, and T. Walsh. (2021). *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*. [Online]. Available: http://ai100.stanford.edu/2021-report

[87] K. Lu, R. Li, X. Chen, Z. Zhao, and H. Zhang, "Reinforcement learning-powered semantic communication via semantic similarity," 2021, *arXiv:2108.12121*.

[88] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.

[89] Y. Wang, M. Chen, W. Saad, T. Luo, S. Cui, and H. V. Poor, "Performance optimization for semantic communications: An attention-based learning approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[90] F. Lotfi, O. Semiari, and W. Saad, "Semantic-aware collaborative deep reinforcement learning over wireless cellular networks," 2021, *arXiv:2111.12064*.

[91] W. J. Yun, B. Lim, S. Jung, Y.-C. Ko, J. Park, J. Kim, and M. Bennis, "Attention-based reinforcement learning for real-time UAV semantic communication," in *Proc. 17th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2021, pp. 1–6.

[92] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," 2018, *arXiv:1803.11485*.

[93] E. Uysal, O. Kaya, S. Baghaee, and H. B. Beytur, "Age of information in practice," 2021, *arXiv:2106.02491*.

[94] H. B. Beytur, S. Baghaee, and E. Uysal, "Towards AoI-aware smart IoT systems," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2020, pp. 353–357.

[95] A. Molin, H. Esen, and K. H. Johansson, "Scheduling networked state estimators based on value of information," *Automatica*, vol. 110, Dec. 2019, Art. no. 108578. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000510981930439X

[96] O. Ayan, M. Vilgelm, M. Klügel, S. Hirche, and W. Kellerer, "Age-of-information vs. Value-of-information scheduling for cellular networked control systems," 2019, *arXiv:1903.05356*.

[97] B. T. Bacinoglu, Y. Sun, and E. Uysal, "On the trackability of stochastic processes based on causal information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2228–2233.

[98] 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 (SSP)*, Jul. 2021, pp. 521–525.

[99] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787

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

[101] B. Wang, X. Deng, W. Liu, L. T. Yang, and H.-C. Chao, "Confident information coverage in sensor networks for field reconstruction," *IEEE Wireless Commun.*, vol. 20, no. 6, pp. 74–81, Dec. 2013.

DYLAN WHEELER (Graduate Student Member, IEEE) received the A.S. degree from the Highland Community College, Highland, KS, USA, in 2016, the B.S. degree in engineering from Ottawa University, Ottawa, KS, USA, in 2018, and the M.S. degree in electrical and computer engineering from Kansas State University, Manhattan, KS, USA, in 2021, where he is currently pursuing the Ph.D. degree with the Cyber-Physical Systems and Wireless Innovations Research Group. His research interests include semantic communications, machine learning, artificial intelligence, and the Internet of Things technologies for beyond-5G wireless networks. He is also a member of the Cyber-Physical Systems and Wireless Innovations Research Group, Kansas State University.

BALASUBRAMANIAM NATARAJAN (Senior Member, IEEE) received the B.E. degree (Hons.) in electrical and electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 1997, the Ph.D. degree in electrical engineering from Colorado State University, Fort Collins, CO, USA, in 2002, and the Ph.D. degree in statistics from Kansas State University, Manhattan, KS, USA, in 2018. He is currently a Clair N. Palmer and Sara M. Palmer Endowed Professor and the Director of the Cyber-Physical Systems and Wireless Innovations Research Group, Kansas State University. His research interests include statistical signal processing, stochastic modeling, optimization, and control theories. He has worked and published extensively on modeling, analysis and networked estimation, and control of smart distribution grids and cyber physical systems in general. He has published more than 200 refereed journals and conference papers. He has served on the editorial board for multiple IEEE journals, including IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.