Title

Design of Interstellar Digital Communication Links: Some Insights from Communication Engineering

Journal

Acta Astronautica, 78(1)

ISSN

0094-5765

Authors

Messerschmitt, David G
Morrison, Ian S

Publication Date

2011-10-24

Peer reviewed

Design of Interstellar Digital Communication Links: Some Insights from Communication Engineering[☆]

David G Messerschmitt[a], Ian S Morrison[b]

[a] *Department of Electrical Engineering and Computer Sciences, 253 Cory Hall, University of California, Berkeley, California 94720-1770, USA, messer@eecs.berkeley.edu*
[b] *Australian Centre for Astrobiology, Biological Sciences Building, University of New South Wales, Kensington, New South Wales, 2052, Australia, ian.s.morrison@student.unsw.edu.au*

Abstract

The design of an end-to-end digital interstellar communication system at radio frequencies is discussed, drawing on the disciplines of digital communication engineering and computer network engineering in terrestrial and near-space applications. One goal is a roadmap to the design of such systems, aimed at future designers of either receivers (SETI) or transmitters (METI). In particular we emphasize the implications arising from the impossibility of coordination between transmitter and receiver prior to a receiver's search for a signal. A system architecture based on layering, as commonly used in network and software design assists in organizing and categorizing the various design issues and identifying dependencies. Implications of impairments introduced in the interstellar medium, such as dispersion, scattering, Doppler, noise, and signal attenuation are discussed. Less fundamental (but nevertheless influential) design issues are the motivations of the transmitter designers and associated resource requirements at both transmitter and receiver. Unreliability is inevitably imposed by non-idealities in the physical communication channel, and this unreliability will have substantial implications for those seeking to convey interstellar messages.

Keywords: SETI, METI, interstellar, digital, communications

1. Introduction

The prospect of interstellar communication has long been a subject of fascination. Some entryways into the extensive literature on this topic can be found in [1, 2, 3, 4].

The prospect of interstellar communication has long been a subject of fascination. Some entryways into the extensive literature on this topic can be found

[☆]Presented at the Second IAA Symposium on 'Searching for Life Signatures', 6-8 October 2010, Kavli Royal Society International Centre, Chicheley Hall, Buckinghamshire, United Kingdom.

in [1, 2, 3, 4]. While the vehicles for conveying messages can include radio and optical radiation, and more exotic means such as neutrinos, gravitational waves or physical artifacts, we focus here on radio frequencies. There are longstanding programs attempting to receive signals [5] (called "SETI" [1]), many of which follow approaches recommended over four decades ago [6]. There have also been messages transmitted (called "METI" [7]), including the "Arecibo message" [8] in 1974 and more recently "Cosmic Call", "Teen Age Message" and "A Message from Earth" [3].

While the vehicles for conveying messages can include radio and optical radiation, and more exotic means such as neutrinos, gravitational waves or physical artifacts, we focus here on radio frequencies. There are longstanding programs attempting to receive signals [5] (called "SETI" [1]), many of which follow approaches recommended over four decades ago [6]. There have also been messages transmitted (called "METI" [7]), including the "Arecibo message" [8] in 1974 and more recently "Cosmic Call", "Teen Age Message" and "A Message from Earth" [3].

Here we adopt a specific perspective, digital communications engineering [9] and its close ally computer network engineering [10]. Sub-disciplines of electrical engineering and the computer sciences, these are concerned with the design of point-to-point, broadcast, and multiple-access systems for communication of information in digital form, as well as multipoint networks built on these, terrestrially as well as in near space (our solar system). Our interest is end-to-end communication systems that convey information, and thus any insights gained from communications engineering are relevant to both SETI (since they inform what to look for) and METI (since they inform what to transmit).

While the engineering community's extensive experience in terrestrial systems is arguably highly relevant to interstellar communications, the interstellar scenario also introduces challenges outside of this experience bank. Thus, we do not claim that communication and network engineering harbor all the answers, and strongly advocate alternative (and even contradictory) ideas as well. For example, while digital communication has relevant advantages (such as the ability to operate at arbitrarily low signal-to-noise ratio and robustness for unreliable and intermittent transmission), it also provides some advantages that are largely irrelevant for interstellar communications (such as source compression and statistical multiplexing of a multiplicity of sources and channels). The recently popularity (and near universality) of digital modulation need not imply that analog modulation (such as persists quite happily in broadcast AM and FM radio) could not present interesting possibilities, and indeed has been one of the key methods of recent experiments [3].

Before exploring specific ideas, some general considerations that the reader should keep in mind are discussed.

1.1. Discovery vs. communication

To establish end-to-end communication, the receiver must first establish the *presence* of a credible signal, and specifically that the signal is of both technological and extraterrestrial origin. We call this the *discovery* stage. Once a

candidate signal has been discovered, the receiver's attention can turn to extracting information from the signal, which we call the *communication* stage. Principally because more is known about the signal (or can be estimated) during the communication phase, discovery is considerably more challenging than communication. Thus, our focus in this paper is on discovery, especially in terms of what a transmitter can do to facilitate discovery.

1.2. Impossibility of coordination

Undoubtably the greatest differentiator between the challenge of interstellar communications and our own experience in terrestrial system design is the impossibility of any form of explicit coordination between the transmitter and receiver designs during the discovery phase. We assert that this lack of coordination can be partially compensated by close attention to the underlying constraints, objectives and principles of communication link design. We call this *implicit design coordination*. There is reason to be confident that the transmitter and receiver, addressing a common set of physical laws, propagation characteristics, and other impairments, will at least arrive at similar conclusions as to the basic elements of a design.

1.3. Anthropocentrism

There is always a danger that we rely too much on our human perspective and experience, which may be singular in some ways. This is called anthropocentrism [11]. A central idea of this paper is to avoid this trap (or at least attempt to avoid it) by basing the system design on "universal" constructs such as our understanding of physical constraints and impairments and mathematical optimization to criteria that are motivated by relevant design objectives (such as minimum energy expenditure or cost). However, the reader is advised to maintain a vigilance for anthropocentrism when judging our approaches.

1.4. Fundamental limits

One possible driver of anthropocentrism is our current technologies, which may be far less capable (or less likely more capable) than those of an extraterrestrial civilization. Subject to (hopefully shrinking) shortcomings in our knowledge of the physical laws we share with such a civilization, we on earth are arriving at an understanding of fundamental physical limitations to both communication [12] and computation [13]. These limitations place boundaries on future technological progress, and as a corollary the assumptions made by another civilization about our technological capabilities. Our technologies today are capable of reaching near the fundamental limits of communication at radio frequencies for a given size of antenna aperture, and have formed the basis for estimates of interstellar communication possibilities [14]. On the other hand, we are remarkably far from approaching the physical limits of computation [13], which suggests there is potential for us to have dramatically enhanced computational resources, either in our own future or as over-optimistically attributed to our present by an extraterrestrial civilization. The primary implication of this

is to discovery, and in particular the feasibility of searching over large parameter sets.

1.5. Motivation

In the absence of explicit coordination, the compatibility of transmitter and receiver designer motivations are influential. The range of possible motivations and their nuances is doubtless a huge topic, but if we narrow to the influence of motivation on transmitter and receiver design we can identify three main categories: *value* (for example benefiting from the knowledge or experience of an extraterrestrial civilization), *curiosity* (for example seeking evidence for the existence of an extraterrestrial civilization, and hence the existence of intelligent life), and *altruistism* (providing value or satisfying curiosity of others without regard to self-benefit). (We omit possible motivations within the category of "hostility".) Among these possibilities, there are limited compatible combinations of transmitter (receiver) motivations, the principal ones being value (value) in a two-way communication, altruism (value) in one-way communication, and altruism (curiosity) in a non-information-bearing beacon.

1.6. Simplicity

During discovery, the receiver has no specific knowledge of the transmit signal and thus must search not only over parameters like time and frequency, but also over a range of hypothesized signal types, and associated parameterizations like bandwidth and power. One of the fundamental (although hardly surprising) insights of communications engineering is that the more you know about the specific waveforms of an incoming signal, the greater the sensitivity[1] of a receiver detecting the presence of that signal, with an obvious boundary when the signal waveform is known exactly. To maximize sensitivity, and also to limit computational complexity, most current SETI efforts focus on a very specific signal, such as an information-devoid single-frequency sinusoid [6].

There is a strong relationship between available computational resources, the range of possibilities for signal types that can be searched, the rate at which unknown parameters like time, frequency, and bandwidth can be searched, and the probability of discovery. The transmitter can increase the probability of discovery by reducing the size of this search space, and one of the primary tools for doing so is by practicing the principle of Occam's razor; that is, by making the signal structure no more complicated than necessitated by design goals, and even artificially relaxing design objectives.[2] This is predicated by assumptions about the computational resources that the receiver brings to bear, which relates to available resources and technology. Thus, "simplicity" is a moving target, but in the interest of expanding the range of civilizations that can discover its signal,

[1] By "sensitivity" we mean the incoming power or energy required to detect the signal with specified objectives for the probability of "detection" and "false alarm".

[2] For example, achieving communication near fundamental limits [6] requires complex techniques that we believe are unlikely to be practical in the absence of explicit coordination.

an altruistic transmitter designer may deliberately make limiting assumptions about receiver resources and capabilities, and this argues in favor of simple over complex.

2. System architecture

As engineers, our foremost inclination is to exploit the "separation of concerns" inherent in modularity [15]. This divide-and-conquer approach decomposes system functionality into pieces with strong internal cohesion but weak coupling [15]. In communications, networking, and software systems, layering [10] is used as a foundation of modularity. We find that a simple form of layering is useful as a conceptual model for the separation of concerns as shown in Figure 1, and the topics in this paper are organized accordingly. Layers build up more complex systems by specializing and exploiting the layers below. Here we address three layers, each with a distinctive grouping of functionality, and each depending on the following less specialized layer:

Message layer. The message creation and interpretation layer (which we call the "message layer") encodes the intended meaning of the message as a sequence of recognizable symbols drawn from a finite alphabet with $M < \infty$ elements. The presumption that the message meaning is represented with a finite alphabet we call the *digital assumption*. The receiver interprets the similar (if not identical) sequence of symbols, extracting the intended meaning.

Reliability layer. The reliability redundancy and assertion layer (which we call the "reliability layer") reflects the intrinsic unreliability of the signal layer as described later. Its purpose is to enable the receiver to extract a more reliable representation of the message sequence. It begins in the transmitter by dividing the message into smaller segments, and then adding protocol elements (non-information-bearing symbols added for several purposes, including delimiting the message segments and assessing their reliability in the receiver), and adding redundancy that allows more reliable segments to be extracted in the receiver (a simple example would be transmitting each segment twice). In the receiver, the message segments are reassembled into a representation of the original message using a process of *assertion*, which means making use of added protocol elements to identify the original message segments and using redundancy to combine or replace segments in ways that attempt to reverse the effects of unreliability. This reassembled message will inevitably harbor inaccuracies relative to the intent of the transmitter because, although redundancy can render mistakes less frequent, it can never eliminate them entirely as described later.

Signal layer. The signal generation and detection layer (which we call the "signal layer") converts the sequence of symbols including the message, redundancy, and protocol elements into a continuous-time radio signal for transmission. The receiver uses statistical detection algorithms to attempt to extract an accurate representation of the transmitted symbols. The extracted symbols will inevitably be an unreliable representation of the transmitter intent due to external impairments like natural sources of noise corrupting the radio signal; major gaps in reception due to the orientation of the earth relative to the line of sight;

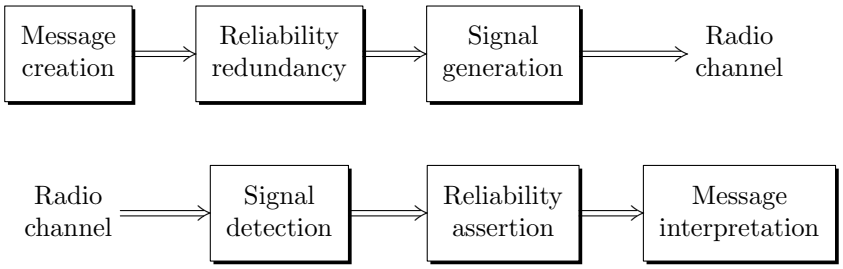

Figure 1: Three distinct groupings of functionality, called layers, in message reception and interpretation, illustrated from the receiver's perspective.

interference of multiple signal paths through the interstellar medium (ISM); periods during which the receiver is inattentive; local sources of radio-frequency (RF) interference; and others.

This paper addresses only the signal and reliability layers; message creation and interpretation is left to others (see Part III of [4]). Our emphasis is on demonstrating the significance of the signal and reliability layers, and sensitizing those addressing the message layer as to the burdens arising from the inevitable departure of the other layers from ideal.

3. Implicit design coordination

How does implicit design coordination actually work? We present some ideas here.

3.1. Optimization as a coordination tool

The portions of Figure 1 that comprise the reliability and signal layers in the transmitter and receiver fall within an end-to-end communication system. Communication system design is, among engineering disciplines, uniquely driven by considerations of mathematical optimality. This is because mathematical models of communication are both accurate and tractable, and as a result most advances arise from theoretical considerations as opposed to experimentation or intuition[3]. This is a fortunate circumstance for implicit coordination. Given compatible assumptions about physical processes and compatible optimality objectives, it is credible that designers of the transmitter and receiver may arrive at compatible conclusions through mathematical optimality.

[3] An exception is that portion of the system that comprises the conversion of the signal from electrical to radio or optical, and the antenna technology associated with the transmission and reception of such radio or optical signals, where physics-based considerations of physical design reign heavily.

Optimization, however, does not completely constrain a design. Where there remains design freedom, in our judgment designers should be guided by simplicity (Section 1.6). We presume that techniques of the scope and complexity commonly used in today's terrestrial communications systems are impractical in the absence of explicit coordination. An example is the three-layer model of Figure 1. While it represents a natural grouping of functionality, and one based on decades of design experience on earth, considerations of optimality have led terrestrial designers to intertwine elements of the reliability and signal layers in complex ways.[4] Thus we advocate keeping these functions entirely separate in our interstellar context.

3.2. Account for the transmitter perspective

Perusal of the SETI literature reveals a preoccupation with design of the receiver. While this is reasonably driven by our own focus on the reception (as opposed to generation) of artificial interstellar radio signals, from the perspective of implicit design coordination this is backwards. It is the transmitter designer who chooses the signal, and therefore considerations of signals to be sought should focus on the knowledge and perspective of the transmitter designer. This does not mean that receiver design considerations are irrelevant, since these need to be considered by a transmitter designer.

3.3. Physical and resource constraints

Constraints imposed by physical propagation of signals through the ISM and by the practicalities of constructing effective transmission/reception equipment may be viewed as distractions, but they help contribute to implicit design coordination by offering guidance to design choices and the range of parameters to search (Section 4).

4. Signal layer

4.1. Signal waveform generation

The signal layer is responsible for the transfer of a sequence of uninterpreted and unreliable data bits[5] from transmitter to receiver. At the transmitter each

[4] As an example, consider error correction coding methods. Historically these have relied on algebraic techniques (like finite fields) that could be applied to the discrete information representations in the reliability layer. However, experience has shown that geometric techniques applied to the "analog" representations in the signal layer are more effective for many applications. A search over the space of possible such mappings we believe to be beyond computational resources even far more advanced than our own.

[5] This is analogous to Layer One (the "physical layer") in the terrestrial open-system interconnection (OSI) end-to-end communication model [16]. Although we conform to the terrestrial assumptions in using "bits", other possible symbol alphabets should be kept in mind.

segment of symbols provided by the reliability layer (including message, redundancy and protocol) is mapped into a waveform suitable for transmission at radio frequencies, a process normally referred to as "modulation"[6].

This may be more involved than a simple mapping. For example, spread-spectrum modulation [19] introduces an additional artificial bandwidth-spreading operation in the interest of interference immunity in the receiver vicinity.[7] It is also possible to build redundancy into the waveforms in the interest of increasing noise immunity by increasing the Euclidean distance between waveforms, a process known as signal-space coding [20]. This is a different form of redundancy than used within the reliability layer as described in Section 5, and is necessary to achieve bit rates approaching fundamental communication limits [21]. These approaches trade increasing complexity for specific performance goals. It is a matter of judgment whether these violations of Occam's razor are justified by the gains, and whether they contribute positively or negatively to implicit coordination. Henceforth the term redundancy is restricted to the generation of redundant symbols (or segments of symbols), as discussed in Section 5.

4.2. Discovery and communication stages

Since we are focused on the discovery stage (Section 1.1), and as discussed further in Section 5, strongly separating the discovery and communication stages so that discovery is accomplished entirely within the signal layer greatly reduces the receiver's search space. Following discovery, many resources can be devoted to reverse engineering the reliability and message layers. This is important because these layers have inherently greater design freedom, are less subject to optimization, and computational requirements are reduced by parameter estimates obtained in discovery.

4.3. Trading transmitter and receiver resources

One basic question, and an opportunity for incompatibility between transmitter and receiver, regards the partitioning of cost and complexity between transmitter and receiver. There is a design tradeoff since more resources devoted to the transmitter (computation, power, antenna area, etc.) can reduce the resources required in the receiver. For example, the transmitter could emit the smallest transmit power consistent with fundamental limits of detection as practiced at a target receiver, or alternatively increase transmit power in the interest of making the receiver designer's job easier. It is common for SETI researchers to make the assumption (whether consciously or not) that the transmitter should bear at least as large a burden as the receiver, and perhaps larger

[6] For multi-target scanning it is favorable to use a high-gain narrow-beam antenna swept across the sky, like a lighthouse beam. In this case, the same segment may be transmitted multiple times in different directions before the next segment is sent. This approach has been suggested most recently in [17, 18] citing cost-efficiency.

[7] In terrestrial systems, spread spectrum offers other advantages such as statistical multiple access that takes natural advantage of fluctuations in data rate. It is difficult to see how these advantages might accrue to interstellar communication.

in the face of our relatively immature technology. A transmitter designer may take the opposite view, with the result that adequate coordination is never achieved.

This is an issue strongly influenced by motivations (Section 1.5). In a one-way communication scenario, one party to the cooperative effort designs only a transmitter, and the other only a receiver. A transmitter designer motivated by altruism may be willing to devote dramatically more resources with the goal of increasing the probability that someone somewhere may derive value or satisfy their curiosity. On the other hand, this alturistic transmitter designer will also observe that the receiver designer is deriving the greater value under this scenario, and therefore may expect the receiver to devote correspondingly greater resources. The latter is the traditional economics perspective. In a two-way communications scenario, there are two parties to the cooperation, each expected to design and construct both a transmitter and a receiver. In this case, both parties are likely to be motivated by value, and both parties expect to incur the cost of both a transmitter and a receiver. The issue then becomes partitioning of resources to minimize total end-to-end cost [17], which is equivalent to minimizing the cost incurred by each party individually. A large speed-of-light latency reduces the value to both parties, reducing the cost that each is willing to incur and thus arguing in favor of limiting attempts for two-way communication to nearby stellar systems. Such a limitation reduces the cost (due to lower impairments and signal attenuation) and increases the value (through smaller latency).

This is an example where the communication and discovery phases deserve independent consideration. Making the signal easier to discover makes it more probable that it will be discovered someplace and sometime. It may even make sense to devote resources to a separate non-information-bearing attractor beacon designed exclusively for easy discovery by virtue of either structure or power level and by virtue of relative simplicity imbued by being non-information bearing. The argument that an unmodulated carrier signal or narrow pulse-like signal is easy to discover is a driver for many of the current SETI searches at radio frequencies [6]. A separate information-bearing signal may be found nearby, making it easier to discover by constraining the search parameter space.

On the other hand, the economic argument in favor of weighting the burden toward the receiver (to the cost benefit of the transmitter) is far stronger for the communication stage. From the transmitter perspective, it is wasteful to devote extraordinary resources when, with high probability, nobody is actually receiving and decoding the messages; that is, no receiver is benefiting and even altruistic urges are not satisfied. It makes more sense to devote extraordinary resources when there is a benefit, and only the receiver can make that determination. Further, the receiver has the facility to expend resources to communication only when it is certain of benefit; that is, following a successful discovery.

These considerations favor a strong separation of discovery and communication, even to the extent of entirely separating the signals. However, it is certainly technically feasible to discover an information-bearing signal, and thus combine

the two functions. This is a normal mode on earth, where there is also the luxury of transmitter-receiver coordination. In light of this, we believe it appropriate to explore the possibility of an information-bearing signal that is also relatively easy to discover [22, 23, 19]. We cite early examples in Section 4.6. This is attractive from several perspectives, including alignment of costs and benefits and Occam's razor, and has profound implications for effective discovery strategies.

4.4. Physical constraints

Interstellar space contains trace amounts of ionized hydrogen at varying densities and in motion with respect to any transmitter or receiver. A radio signal propagating through the ISM will experience both frequency- and time-dependent changes in propagation characteristics, resulting in several forms of signal degradation. The principal forms are *dispersion* [24] and *scattering* [25].

Dispersion introduces a frequency-dependent group delay, with larger delays at lower frequencies. For non-monochromatic signals this results in delay-spread, which distorts the shape of the modulated symbols and may cause intersymbol interference (ISI) [20]. Dispersion can be reversed by phase equalization in the receiver if the dispersion measure (DM), a measure of electron density in the ISM, is known. Before discovery the DM will be unknown. It varies depending on the line of sight, and while subject to astronomical observations some uncertainty always remains. Either the transmitter chooses a signal design and parameters with tolerance to dispersion uncertainty [24] (e.g. higher carrier frequency or lower bandwidth) or the receiver must accept the computational burdens of estimating or searching over the DM. ISI can be reduced by making data symbols longer and/or spread further apart, resulting in lower data rates. These considerations offer helpful guidance to both transmitter and receiver as to a beneficial signal design and parameterization.

Scattering is the result of destructive or constructive interference among signal components traversing different paths through the ISM [25]. It results from inhomogeneities in the electron density and resulting refractive and diffractive effects. Analogous to multi-path fading in terrestrial wireless communications systems [26], scattering results in time-varying amplitude/phase variations across the signal bandwidth, introducing nulls and peaks at certain frequencies that typically change with time. Increasing the signal bandwidth such that the nulls represent only a small fraction of the total signal bandwidth counteracts scattering, but also exacerbates the effects of dispersion. *Frequency diversity* transmits the same signal at multiple carrier frequencies such that it is statistically unlikely that all carriers will experience a null at the same time. Combining the signals at the receiver will result in an averaging effect and provide a more consistent signal strength. The difficulty with diversity is that, prior to discovery, the receiver does not know which different carrier frequencies to combine.

The relative motions of the transmitter, ISM and receiver gives rise to Doppler effects. Specifically, the acceleration component of relative motion causes a quadratic time warping of the signal. While this is insignificant over short time intervals, over longer intervals it may be problematic for signal detection. However, it can be compensated at the transmitter and receiver, each

of whom is aware of the component of their own acceleration along the line of sight. To accomplish this, however, transmitter and receiver must agree on a common inertial frame, such as the center of the Milky Way. The argument for such compensation is compelling at both transmitter and receiver.

Due to the limited time-coherence of the ISM channel, there may be advantages to limiting the timespan of transmitted waveform bursts such that the ISM propagation characteristics are relatively constant throughout the duration of each burst. However, shortening the burst duration can have an adverse effect on reliability, because it reduces the dimensionality of the signal and limits the gains achievable through signal-space coding. This effect may be overcome through the use of time diversity techniques implemented within the reliability layer, such as spreading across multiple waveform bursts the influence of the symbols from any given message segment. However, time-diversity techniques are subject to the fundamental trade-off that exists between latency and reliability (as described in Section 5.3). This may therefore lead us to view *time* as a physical constraint that provides a degree of implicit coordination.

4.5. Resource constraints

Like physical constraints, resource constraints provide multiple forms of guidance as to design choices, and therefore provide some implicit coordination between transmitter and receiver.

The inverse-square law for receive power as a function of distance necessitates high transmitter power levels even when the receiver operates at or near fundamental limits. Even for civilizations more advanced than our own, we can reasonably presume that energy resources are not unlimited, and that there are competing usages for those finite energy resources. As argued in Section 4.3 motivation is crucial, as an altruistic transmitter in a one-way communication scenario is more willing to devote resources (such as transmit power greater than necessary) than is a value-conscious transmitter/receiver designer in a two-way communication scenario.

Once the transmitter designer has chosen a radiated power level, there is still an opportunity to minimize total cost at long transmission ranges by splitting the cost between the RF power source and antenna [17]. A large high-gain antenna increases the range for a given RF input power. Cost efficiency considerations therefore argue for high antenna gains and scanning the sky with finite dwell and revisit times on any given target. This results in the receiver observing the signal from an individual transmitter as a transient source, a reason that there is an increasing focus on transient signals in SETI.

In choosing its transmit power, the transmitter designer must make assumptions about the antenna area (and hence antenna gain) in the receiver. Since this is related to the cost of the receiver, here again motivation is determinant. It is conceivable that a transmitter has assumed an antenna gain in the receiver that is many orders of magnitude larger than any yet built on Earth. Indeed, this is one possible explanation for the absence of signal discoveries to date.

Objectives and assumptions must also be set with regard to radio-frequency interference (RFI) at both transmitter and receiver locations. The transmitter

designer must also make assumptions as to the RF interference environment in the vicinity of the receiver. The receiver designer has limited control over this environmental factor, although the receive antenna can be moved to a "quiet" location in space or on another astronomical body at increased expense. This assumption influences both transmit power level and signal design [19], and again there is a tradeoff with transmitter cost and the resources devoted to discovery at the receiver that is influenced by motivations (Section 1.5). In the absence of knowledge of the receiver's local interference environment, it would seem prudent for the transmitter to assume larger levels of interference, even if such levels of interference are not typical for their own local environment. This suggests the favoring of signals with large time-bandwidth product [19], and in view of the time-varying propagation effects this in turn suggests higher-bandwidth signals.

The choice of carrier frequency is also influenced by several design factors. For a given transmitter antenna area and power, the maximum propagation distance increases as carrier frequency increases due to increasing antenna gain. This suggests that it is advantageous to use the higher end of the microwave window, e.g. towards 10 GHz.[8] As it is technologically easier and cheaper to achieve a bandwidth that is a small fraction of the carrier frequency, this also loosens the constraints on signal bandwidth. Propagation effects also favor high carrier frequency, as both dispersion and scattering effects decrease rapidly with carrier frequency.

In contrast to terrestrial systems, which must share spectrum across many competing uses, the bandwidth of the ISM is not inherently limited by multiple access objectives. Due to the great distances involved, highly directive beams are at minimum very desirable (if not a necessity) to limit the transmit power, and thus even if there are multiple communications ongoing they are not likely to interfere even if they share a common spectrum. More likely to be a concern to the transmitter is a desire not to interfere with the radio astronomy activities of receiving civilizations. A high-powered narrowband beacon that concentrates all its power at one frequency could be said to resemble a 'jammer'. In this regard spread-spectrum signaling may be favoured as a means of reducing the peak power spectral density. When optimum detection techniques like matched filtering are employed during the communication stage, increasing bandwidth does not adversely affect noise immunity. Approaches that trade bandwidth for power efficiency are also attractive, as illustrated by Section 4.6. On the other hand, increasing bandwidth does make discovery more challenging [24] due to imprecise knowledge of dispersion and scattering, and higher bandwidth involves some technological challenges in both transmitter and receiver.

[8] If the transmitter assumes the receiver is located in space or elsewhere without an absorptive atmosphere, even higher carrier frequencies could be attractive.

4.6. Example

Information-bearing signals can be designed for detectability, such that they can be discovered without the necessity of a separate attractor beacon. While this is routine in terrestrial systems, it is more challenging for interstellar communication because of the lack of explicit coordination. What is needed is a property of the signal that distinguishes it from not only natural sources of noise, but unknown or varying levels of such noise [22], as well as a receiver algorithm that highlights that property even in the presence of dispersion and scattering. Ideas currently being explored include superimposing a delayed signal replica that triggers an autocorrelation peak in the receiver [22]; detection of individual symbols by matched filtering [24]; and autocorrelation between successive symbols over a range of assumed symbol period estimates [23].

This last technique, *symbol-wise autocorrelation* (SWAC), illustrates a 'self-discoverable' information-bearing wideband signal. Assume the received data waveform is spread-spectrum binary phase-shift keying (BPSK) with a symbol rate of 2 symbol/s. Each symbol is actually a complex waveform consisting of smaller "chips" at a rate of 1000 chips per symbol. This is the "spread-spectrum" part, and its purpose is to give more robust interference rejection in the receiver [19]. Figure 2(a) illustrates the frequency spectrum of this signal embedded in noise, where the level of the signal is lower than that of the noise. It is difficult to ascertain the presence of a signal from the spectrum. However, this signal (and many other information-bearing waveforms like it) exhibits a property called cyclostationarity [27] that can be observed by an autocorrelation algorithm in the receiver. SWAC is a variant of autocorrelation algorithm that is particularly effective in detecting randomly modulated spread-spectrum signals of the type pictured in Figure 2(a). Applying the SWAC algorithm to this signal produces an autocorrelation peak at the symbol period (500 ms), as illustrated in Figure 2(b). Not knowing the symbol rate, the receiver has had to perform an autocorrelation at many different candidate lags, but one clearly stands out. The sensitivity of this type of detector is below that of a matched filter, but it has the advantage that it requires no prior knowledge of the carrier frequency, signal bandwidth, modulation method or symbol alphabet [23].

This example illustrates a "distinguishing feature" of the information bearing signal, a feature that is triggered by a specific algorithm executed in the receiver [22]. While SWAC works for a variety of cyclostationary signal types, there are other signal types requiring other techniques for effective detection. The receiver needs to guess the right algorithm, or apply multiple algorithms from a set that includes the one appropriate for the signal type chosen by the transmitter.

5. Reliability layer

The purpose of the coordinated reliability layer in transmitter and receiver is to present to the receiver's message interpretation layer the most reliable possible representation of a sequence of symbols representing the message. An

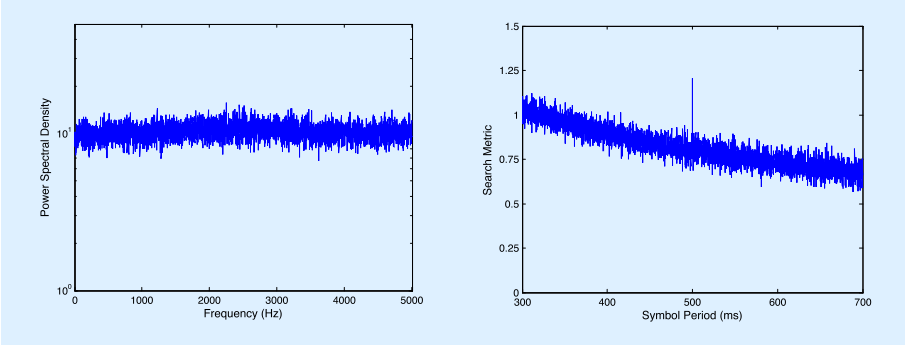

Figure 2: An illustration of symbol-wise autocorrelation (SWAC). (a) On the left is the power spectrum of a binary-antipodal spread-spectrum modulated data waveform in noise. (b) On the right is the symbol-wise autocorrelation of this waveform calculated for different hypothesized symbol periods. The observed peak corresponds to the correct symbol period.

important design constraint should be that discovery is allowed to proceed successfully within the receiver's part of the signal layer, implying that to accomplish detection the receiver requires no knowledge of the reliability protocols assumed by the transmitter, nor of the message representation. Once discovery has occurred, the transmitter can reasonably assume that the receiver is willing to devote extraordinary time and resources to "reverse engineering" of the reliability protocols and message content/intent. This is fortunate, because the reliability layer will require relatively complex protocols to perform its function, and there is considerable design freedom, implying that a search over different reliability protocols during discovery is likely to be prohibitive in computational requirements (at least for a civilization at our stage of development) and very likely fruitless.

5.1. Functions of the reliability layer

The reliability layer includes three generic functions listed in Table 1. Each of these functions is associated with a specific type of protocol element that will be embedded within the stream of symbols passed to and from the signal layer.

5.2. Implications for message interpretation

The receiver designer should expect the sequence of symbols passed from the signal layer to not only represent a message, but also to embed protocol elements such as those listed in Table 1.

Even with these added elements, the reliability layer cannot achieve perfect reliability. To even begin to approach perfect reliability requires two-way protocols of the type used in the Internet's TCP protocol, in which correctly received message segments are acknowledged by a message sent on a return channel from receiver to transmitter. Two-way protocols are not feasible for an interstellar

Table 1: Generic functions achieved by implicitly coordinated reliability layers in the transmitter and receiver and associated types of protocol elements embedded in the stream of symbols.

Function	Protocol element	
Partitioning	Delimiter	The boundary between message segments may be identified by a delimiter, such as a fixed sequence guaranteed not to fall within the segment content. Fixed-length message segments is a simplification.
Sequencing	Sequence number	Each message segment may have a sequence number identifying its position in the sequence. This is particularly valuable in case of redundant transmission of the same segment at different times.
Correctness	Check sum	A message segment, including sequence number, may include an element that allows the receiver to detect corruption of that segment. Usually this element is the result of a mathematical function calculated on the remainder of the segment. A simple example is an arithmetic sum of symbols.
Recovery	Redundancy	A corrupted segment can sometimes be recovered if it is accompanied by redundancy. A simple example would be the retransmission of the segment at three disparate times, with the receiver assuming that two identical segments are a correct representation of the transmitter intent.

channel wherein the communication latencies are on the order of tens or hundreds or even thousands of years. Absent such an acknowledgment mechanism, the transmitter can never know whether a particular message segment has been correctly reconstructed by a given receiver. Absent such knowledge, infinite resources (in terms of, for example, repeated retransmission in perpetuity) would be required to approach perfect reliability.

All the elements in Table 1 reduce the likelihood of message corruption in the receiver, but none of them is foolproof. For example, random errors within a segment can, with some small but non-zero probability, preserve the coherence between message content and the correctness function.

Thus the message interpretation layer in the receiver must be designed with the assumption the message is represented by a sequence of message segments, and that those segments will inevitably suffer from different types of degradation as identified in Table 2.

5.3. Latency in message interpretation

There is a direct trade-off between reliability and latency (time elapsed before complete decoding of the message). If, for example, it is acceptable to increase latency from years to decades or centuries, increasing reliability can be

Table 2: Generic types of degradations that may be suffered by message segments passed from the reliability layer to the message layer in the receiver.

Blind corruption	A message segment may have one or more symbols that are different from the transmitter message.
Identified corruption	A message segment may be tagged as likely to contain one or more symbols that are different from the transmitter message. The message interpreter is free to choose whether or not to discard this segment entirely.
Identified omission	One or more message segments may be missing from the sequence, and identified as such by the reliability layer.
Blind omission	One or more message segments may be omitted from the sequence and not identified as such by the reliability layer. This is less likely than a known omission.
Blind reordering	Two or more message segments may be out of their original order in the sequence.

achieved. For a given latency objective, there is also a tradeoff between the total message length and reliability, as the total communication resource discussed in Section 4.5 must be partitioned between information, information redundancy, and protocol elements. Increasing redundancy (for example sending each message segment three times rather than twice) results in improved reliability but implies a reduction in message length.

5.4. Implications for message creation

Message creation can and should take account of the intrinsic unreliability of the message representation in the receiver. This is largely an issue of being sensitive to the extent and length of error propagation effects in message interpretation. For example, a one-time definition of a particular lexical element (vocabulary word) should be avoided if that element is crucial to interpretation of the entire remainder of the message. Rather, lexical or semantic redundancy in the message will contribute to limiting the impact of message corruption as well as make the message more likely to be interpreted correctly at the semantic level.

6. Conclusions

The paper draws attention to many of the factors influencing and constraining the design of transmitters and receivers for interstellar messaging, these factors flowing from both physical phenomena in the Universe and engineering design considerations related to motivations, objectives, and resource constraints. The central ideas include avoiding anthropocentrism and achieving implicit coordination through basing design decisions on physical principles and

mathematical optimization, and through simplicity of design. Influencing design choices by assumptions about motivations and resources is unavoidable, but we advocate making those assumptions transparent as well as explicitly and comprehensively exploring the implications of alternative assumptions.

We have incorporated the engineering concept of modularity through a layered architecture as a way of identifying and categorizing the needed functions of an interstellar communications system, and as a means of separating their concerns and identifying their inevitable dependencies. Specifically we chose a modularity that separates the functions of signal, reliability, and message, but this may be generalized and elaborated in the future as system design is considered in more detail. In engineering practice, this elaboration typically takes the form of further hierarchical decomposition of the system functionality.

We have emphasized the important distinction between discovery and communication, and advocated a careful modular separation of these functions. The important and unavoidable dependencies we have identified include a tradeoff between transmitter and receiver resources, and a coupling between reliability, throughput, and latency. It is advantageous for discovery to be accomplished entirely within the signal layer, but it likely cannot be divorced entirely from the reliability and message layers as they introduce design considerations that influence the signal layer.

We have emphasized that the constraints and limitations of the reliability layer (which largely flow from the realities of the physical environment, such as noise and latency) have profound implications for message construction and interpretation.

A roadmap for more thorough study of end-to-end digital interstellar communications has been proposed. We have argued that this systematic study can better inform the types of signals to seek in SETI programs, and also inform the design of signals for METI programs. While such a study would draw primarily on the principles and experience of communication engineering, it would be informed by an understanding of the underlying physical processes developed by astrophysicists and astronomical observations. It may also identify gaps in knowledge of the relevant physical processes that can beneficially be addressed by astrophysicists and astronomers.

References

[1] J. Tarter, "The search for extraterrestrial intelligence (SETI)," *Annual Review of Astronomy and Astrophysics*, vol. 39, no. 1, pp. 511–548, 2001.

[2] A. L. Zaitsev, "METI: Messaging to extraterrestrial intelligence," *Searching for Extraterrestrial Intelligence*, pp. 399–428, 2011.

[3] A. L. Zaitsev, "Sending and searching for interstellar messages," *Acta Astronautica*, vol. 63, no. 5-6, pp. 614–617, 2008.

[4] D. A. Vakoch, *Communication with Extraterrestrial Intelligence*. State University of New York Press, 2011.

[5] J. Welch, D. Backer, L. Blitz, D. C. J. Bock, G. C. Bower, C. Cheng, S. Croft, M. Dexter, G. Engargiola, and E. Fields, "The allen telescope array: The first widefield, panchromatic, snapshot radio camera for radio astronomy and seti," *Proceedings of the IEEE*, vol. 97, no. 8, pp. 1438–1447, 2009.

[6] "Project cyclops: A design study of a system for detecting extraterrestrial intelligent life," tech. rep., Stanford/NASA/Ames Research Center Summer Faculty Program in Engineering Systems Design, 1971. Accessed at http://www.archive.org/details/projectcyclopsde00stan on 18 Feb. 2010.

[7] A. L. Zaitsev, "The first musical interstellar radio message," *Journal of Communications Technology and Electronics*, vol. 53, no. 9, pp. 1107–1113, 2008.

[8] D. Goldsmith and T. C. Owen, *The search for life in the universe*. Univ Science Books, 2001.

[9] J. G. Proakis, M. Salehi, N. Zhou, and X. Li, *Communication systems engineering*, vol. 10. Prentice Hall, 2002.

[10] J. F. Kurose, K. W. Ross, and B. Anand, *Computer networking: a top-down approach*. Pearson/Addison Wesley, 2008.

[11] V. Ascheri, "A methodological approach to communication with extraterrestrials," in *Bioastronomy 99*, vol. 213, p. 603, 2000.

[12] T. M. Cover and J. A. Thomas, *Elements of information theory*, vol. 6. Wiley Online Library, 1991.

[13] S. Lloyd, "Ultimate physical limits to computation," *Nature*, vol. 406, pp. 1047–1054, 2000.

[14] S. Shostak, "Limits on interstellar messages," *Acta Astronautica*, 2009.

[15] D. L. Parnas, "On the criteria to be used in decomposing systems into modules," *Communications of the ACM*, vol. 15, no. 12, pp. 1053–1058, 1972.

[16] H. Zimmermann, "OSI reference model–the ISO model of architecture for open systems interconnection," *Communications, IEEE Transactions on*, vol. 28, no. 4, pp. 425–432, 2002.

[17] J. Benford, G. Benford, and D. Benford, "Messaging with cost-optimized interstellar beacons," *Astrobiology*, vol. 10, no. 5, pp. 475–490, 2010.

[18] G. Benford, J. Benford, and D. Benford, "Searching for cost-optimized interstellar beacons," *Astrobiology*, vol. 10, no. 5, pp. 491–498, 2010.

[19] D. G. Messerschmitt, "The argument for spread spectrum in interstellar messaging," draft awaiting publication available at `www.eecs.berkeley.edu/~messer`.

[20] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital communication*. Boston: Kluwer Academic Publishers, 3rd ed., 2004.

[21] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, p. 379423, 1948.

[22] G. R. Harp, R. F. Ackermann, S. K. Blair, J. Arbunich, P. R. Backus, and J. Tarter, *A new class of SETI beacons that contain information.* Communication with Extraterrestrial Intelligence, State University of New York Press, 2011.

[23] I. S. Morrison, "Detection of antipodal signalling and its application to wideband SETI," 2011. this issue.

[24] D. G. Messerschmitt, "Interstellar spread-spectrum communication: Receiver design for plasma dispersion," draft awaiting publication available at `www.eecs.berkeley.edu/~messer`.

[25] S. K. Blair, D. G. Messerschmitt, J. Tarter, and G. R. Harp, *The Effects of the Ionized Interstellar Medium on Broadband Signals of Extraterrestrial Origin.* Communication with Extraterrestrial Intelligence, State University of New York Press, 2011.

[26] D. Tse and P. Viswanath, *Fundamentals of wireless communication.* Cambridge Univ Pr, 2005.

[27] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: Half a century of research," *Signal Processing*, vol. 86, no. 4, pp. 639–697, 2006.