**The Global Catastrophic Risks of**

**Finding Extraterrestrial AI Message**

 **During SETI**

Alexey Turchin

alexeiturchin@gmail.com

Foundation Science for Life Extension, Russia, Moscow, Prospect Mira 124-15

**Abstract**: This article examines risks associated with the program of passive search for alien signals (search for extraterrestrial intelligence, or SETI) connected with the possibility of finding of an alien transmission which includes a description of an AI system aimed at self-replication (SETI-attack). A scenario of potential vulnerability is proposed, as well as the reasons why the proportion of dangerous to harmless signals may be high. The article identifies the necessary conditions for the feasibility and effectiveness of the SETI-attack: ETI existence, possibility of AI, small size of the Seed AI, slow speed of physical interstellar travel, and large distance between civilizations. Needed additions to the SETI protocol are considered: keep the signal’s existence, content and source location secret, don’t run alien programs, jam dangerous signals, wait until the creation of humanity’s own AI. In addition, scenarios in which it may be reasonable to start an alien AI (if it has been downloaded) are explored.

**Keywords**: SETI – METI – existential risk – artificial intelligence – radio transmission – extraterrestrial intelligence

**Highlights**:

* A potentially dangerous SETI-message could consist of the description of a computer, a program for that computer, and a bait.
* The alien program could be a Seed AI, which could convert Earth to a new signal transmitter.
* Viral messages are probably the dominant type of SETI-messages because of the natural selection effect.
* SETI-attack outperforms physical travel in intergalactic distances.
* SETI protocol should be updated to reflect the risks of a SETI-attack.

## 1. Introduction

The idea that passive searches for extraterrestrial intelligence (SETI) have the potential to be dangerous is not new. Physicist Fred Hoyle famously suggested a scheme of alien attack through SETI signals (hereafter called “SETI-attack”) in the novel “A for Andromeda” [1]. In the plot of that work, astronomers receive an alien signal that contains a description of a computer, and a computer program to be run on it. This machine creates a description of genetic code, which leads to the creation of an intelligent creature—a girl dubbed Andromeda—who, working together with the computer, creates advanced technology for the military. The initial suspicion about the intent of the aliens is overcome by the greed for the technology they can provide. However, the main characters realize that the computer acts in a manner hostile to human civilization, so they destroy it, killing the girl.

In the book "Mind Children", Hans Moravec offers a similar scenario of vulnerability: the potential for humans to download a computer program from space via SETI that has artificial intelligence [2]. This AI may promise new opportunities for the owner, and after fooling the human host, is capable of self-replicating millions of copies, destroying the human host, and finally using the resources of the secured planet to send its ‘child’ copies to multiple planets that constitute its future prey. Such a strategy is similar to the strategy of a virus or a digger wasp—horrible, but plausible.

Carrigan wrote about the risk which could be posed by unfiltered signals from space which could be loaded on millions of insecure computers that took part in the *SETI-at-home* program [3]. This proposal was met with a great deal of criticism from programmers who pointed out that first, data and programs are in separated regions of memory in the computers, and secondly, the programming languages in which programs are written are unique products of human technology, such that it is impossible to guess their structure in order to hack them blindly without prior knowledge.

Following that response, Carrigan published a second article, "Do potential SETI signals need to be decontaminated?" [4]. In it, he pointed to the ease of transferring gigabytes of data at interstellar distances, and indicated that such interstellar signals may contain some kind of bait that will encourage people to construct a dangerous device, according to the designs that these signals contain. However, Carrigan didn’t give up his belief in the possibility that an alien virus could directly infect Earth’s computers without human ‘translation’ assistance. While this seems implausible, human help is easily reachable, as will be shown below.

Some other influential works of fiction have also explored the idea of SETI risk. Vernor Vinge’s novel “A Fire Upon the Deep” shows the risk of alien AI found as a result of space archeology [5]. In Carl Sagan’s “Contact”, aliens send a description of a mechanism—not AI, but a worm-hole transporter. In Sagan’s story, the characters discussed possible risks, but decided that aliens could find simpler ways to destroy life on Earth [6]. In Stanisław Lem’s “His Master’s Voice” astronomers find a message that appears to have some form of intelligence encoded within it, found to exist primordially as some universe’s genetic code [7].

Hawking has expressed concern about the risks of contact with aliens [8], and Baum has explored various potentially dangerous scenarios of communication with extraterrestrial intelligence [9]. Michaud has discussed potential dangers of contacts with aliens [10]. Penny suggested that listening for alien signals may be dangerous, but has stated that as the probability of risk can’t be determined, there is no reason not to listen [11]. Hippke asserted that complex messages can’t be decontaminated with certainty [12]. Yampolskiy included alien AI obtained via SETI as one of the possible subtypes of dangerous AI [13].

Yudkowsky’s article, "AI as a positive and a negative factor of global risk", introduced his notion of “Seed AI” [14], or embryonic AI—a minimum program capable of learning and of runaway self-improvement with an unchanged primary goal. He also stated that Seed AI can evolve extremely quickly, and it could easily outsmart humans and take over the world. He said that the risks of self-improving AI are underestimated, so it is impossible to keep AI in a "box"; and that it is impossible to recognize dangerous intentions of an AI in advance. Nick Bostrom outlined a similar set of risks relating to AI in his book “Superintelligence” [15].

Despite many warnings, both fictional and non-fictional, about its potential risks, SETI research continues unabated [16]. New waves of funding continue for this potentially attractive idea. At the same time, SETI possibilities are growing, as the quality of astronomical instruments, and the computer power available to analyze astronomical data improve. These advances mean that humanity will be able to find alien civilizations from further away—across larger distances—if they are there to be found. The number of stars checked for alien signals is growing exponentially, close to millions in the 2010s—and the probability of “success” is also growing at the same rate [17].

The number of projects searching for radio signals of alien origin is growing, partially because of contributions from private individuals. Paul Allen created the Allen Telescope Array for SETI search, which aimed to examine 1 million stars, and to search for other signs of alien activity, like laser flashes [18]. Russian billionaire Yuri Milner donated $100 million for SETI research in 2015 [19].

The interest in SETI has grown since 2016, when the star KIC 8462852 exhibited strange patterns in its light signature, as if being irregularly overshadowed by some other objects. Some think it may be elements of a Dyson sphere built by aliens, or perhaps a beacon to attract humanity’s attention [20]. The star is 1600 light-years from the Earth, which seems too close for SETI-attack (see section 6).

The distance over which a SETI search can be conducted, and thus the number of stars it can reach, is growing. But more importantly, from point of view of a potential SETI attack, is the search, not for complex radio transmissions, but for beacons, which could be found by astrophysical instruments, not directly aimed at searching for alien life. Several beacon-candidates have been found in recent years via astrophysical research, including fast radio bursters, and dimming stars with complex patterns, like KIC 8462852.

Most participants in SETI dismiss any risks relating to passive SETI, as well as active SETI [21]. In a personal communication to the author, Seth Shostak confirmed his awareness of the potential risks of passive SETI but stated that he doesn’t regard it as serious. The risks of passive SETI have received much less attention than the risks of active SETI—messaging extraterrestrial intelligence (METI)—that could make our existence known to aliens [17,22–25].

But, in fact, passive SETI is more dangerous than METI, as the risks are immediate, they could come from a further distance, which includes many more stars and civilizations, and because SETI-attack will be naturally selected to be most malicious (more below).

The goal of this article is to estimate the risk of SETI-attack and suggest possible measures for its prevention. First, necessary conditions for SETI-attack to occur will be identified in section 2, then a plausible scenario of attack will be explored in detail in section 3, in section 4 will be presented an overview of the possible goals of the senders, section 5 is devoted to the model of the SETI-attack depending on the civilization distribution in space, section 6 analyzes possible prevention measures and some safety recommendations for SETI protocols, and section 8 looks at a situation in which starting alien AI would be a preferable outcome.

## 2. Main premises upon which the probability of SETI-attack is based

The possibility of the SETI-attack is based on two main premises, which are necessary conditions for such an attack be technically possible.

1. *Strong artificial intelligence is possible as a classical computer program of finite size,*and this size is small enough that such a program can be transmitted by radio or other means of communication. Even this requirement is too strong—alien AI does not require full artificial intelligence, with consciousness, creativity, and the capacity to pass the Turing test, to be harmful. It just has to be able to win in several games against humans, and have superhuman expertise in the fields of astrophysics, electronics, computer science and game theory. Contemporary computer programs routinely win in narrow fields like chess [26], so the ability for such requirements to be met seems probable. Thus, alien AI could be characterized as a polymorphic virus capable of machine learning.
2. *Technological extraterrestrial civilizations exist in the visible universe*and they are close enough to send large amounts of data, but far enough to make it difficult for them to travel physically to Earth. SETI research is based on a non-negligible probability that alien civilizations exist. Even if the “Rare Earth”—the hypothesis that Earth is the only habitable planet in the visible Universe here[27]—is most probable, the small probability of other civilizations existing dominates the risk analysis landscape. The condition will be deeply analyzed in section 7, where we explore the distribution of civilizations in the universe favoring SETI-attack.

There are several subconditions, one of which is that size of the possible seed AI will fit into data sending technologies. Carrigan showed that size of human minds is of order of magnitude of 1 Gigabyte, and that this amount of data can be sent over interstellar distances using radio transmissions [4].

But large amounts of data could be sent intergalactic distances via large-scale astro-engineering, which will enable extragalactic SETI [28].

 A drawing using a galactic plane of a remote galaxy as its canvas, and Dyson spheres as pixels could send around 1 gigabyte of data over billions of light-years, assuming that one billion stars are converted to Dyson spheres, and that the alien sender finds ways to compensate for distortion due to galactic rotation. While such repeated sending in all directions is a very energy-intensive task, it could be simplified if the attention of the receiver civilization can be attracted by some kind of lighthouse or beacon.

## **3. Algorithm of SETI-attack**

### 3.1 Overview of the SETI-attack structure

One possible scenario of SETI-attack will be explored in this section, though other scenarios are also possible.

A message that is part of a SETI-attack will probably include the following elements:

* A beacon, which attracts attention to the main message
* An introduction to the alien language
* Bait, an explanation of why we should create an alien computer
* Blueprints for a simple computer
* The code for a program to be run on this computer

From this list, the program to be run on the computer is the most important part, and the only one required for a SETI-attack. The other components are needed to explain that it *is* a program, convince us we should run it, and tell us what hardware is required to run it.

### 3.2 The simplest possible computers are universal

The explanation about the computer needed to run the alien program will be as simple and short as possible, which requires simple programming languages. The message could have two levels, where the first part is a just description of a loader program, which will create an optimized computer or complier which will run code more effectively.

Humans have created very simple self-evident computers and programs. For example, the simplest computer, the Turing machine, has only six commands [29]. Such simple computers are based on mathematical universalities, like the Pythagorean theorem, and there are only a few possible variants of the Turing machine [30]. Even the Babbage mechanical machine, designed in the early 19th century, was Turing-compatible [31]. If code for such a computer was received, it would be easy to recognize, and to run it on a virtual Turing machine.

In fact, a Turing machine derivate called the “Wang machine” has only four commands, and seems to be the simplest possible computer [32]. Humans also created self-evident program languages for METI, the specially-designed *CosmicOS*, which has only four symbols. “*CosmicOS* is a self-contained message designed to be understood primarily by treating it as a [computer program](https://en.wikipedia.org/wiki/Computer_program) and executing it” [33].

A Turing machine has only six commands and a tape data recorder. Its full electric scheme will contain only a few tens of transistors or logic elements; thus, it is not difficult to send blueprints of a Turing machine. And even without such blueprints, the code for a Turing machine will be easily recognizable. Turing machines are so simple that one could easily be built from Lego (and has been) [34].

### 3.3 One possible scenario for a SETI-attack

In this subsection one possible scenario for a SETI-attack will be reviewed in detail.

To start, let’s assume that there is an extraterrestrial civilization, with the *intent* to send an AI-empowered message that will enable it to obtain power over Earth. In the next section, we will consider how realistic it is that another civilization would *want* to send such a message.

In order to prove that a vulnerability exists, it is enough to find just one hole in security. However, in order to prove safety, one must remove every potential hole. The complexity of these tasks varies on many orders of magnitude, and is well known to the experts in computer security [35]. This distinction has led to the fact that almost all computer systems, from *Enigma* to the *iPhone*, have been broken. In this section one vulnerability of a SETI program will be demonstrated; however, any error in the thesis doesn’t automatically prove safety of SETI program: other risks are still possible.

The detailed stages of a SETI-attack could be as follows:

1. The sender civilization *creates a signal beacon or lighthouse* in space. The signal this beacon sends clearly shows that its origin is artificial; for example, a star with a Dyson sphere. This beacon can be seen across a distance of thousands of light-years or maybe even hundreds of millions of light-years. There are possible various types of beacon, but the important fact is that its signal could be viewed across long distances, and be easily recognized as of artificial origin. The goal of the beacon is to attract the attention of astronomers in the receiving civilization, so its signal should be obviously atypical of natural phenomena.

2. Near this beacon is *located a radio transmitter with a much weaker signal*, but is information-rich. The lighthouse draws attention to this second radio source. This is exactly what happened after the strange behavior of the star KIC 8462852 was discovered: it was checked for radio signals [36]. This secondary source produces a stream of binary information (i.e., sequences of 0 and 1). The most obvious way (understandable by the recipient) to reduce noise is simple repetition of the signal in a circle. It should be done in reasonable time, probably less than a year, so the signal should convey 100 bits a second to send around 1 gigabit in a year.

3. The simplest way to convey meaningful information to an unknown recipient using a binary signal is by *sending images*. Eye structures appeared independently 50-100 times in the biological history of the Earth [37]; this suggests the presentation of a three-dimensional world with the help of 2D images is probably universal, and is almost certainly understandable to all creatures that can build a radio receiver. That is why we send our messages to aliens mostly in form of images, like the plates on the Voyager probes [38].

Two-dimensional images are not difficult to encode as binary signals: the system which used in the first TVs, line-by-line scanning, would suffice. At the end of each line repeated signal is used to mark the end. Finally, at the end of each frame is placed another signal indicating the end of the frame. This method can be used to send still images or a continuous video.

The mediocrity principle is also at work here: humanity can assume that we are a typical civilization, and the senders are probably aware of the most abundant types of civilization. This is especially true if the attacker is a large interstellar civilisation with many previous successful SETI-attacks (which will be shown later is the most probable profile for attackers). Thus, they send such messages that most typical civilisations will easily understand.

5. A sender civilization should be extremely interested that humanity understand their signals. On the other hand, human scientists will probably exhibit high desire to decrypt the signal. Therefore, there is no doubt that *the pictures will be recognized.*

6. Images and movies can convey a lot of information; they can even train us in learning alien language and show us their world. It is obvious that many can argue about how such films will be understandable. If a certain civilization sends radio signals, and the other receives them, *they should have some shared knowledge*. Namely, they know radio technology: that is, they know transistors, capacitors, and resistors. – as they are needed to build a transmitter. These radio components are quite typical, so that they can be easily recognized in the photographs.

7. By sending *photos depicting radio parts* on the right side of the image, and their symbols on the left, it is easy to convey a set of signs indicating elements of electrical circuits: transistors, capacitors, and resistors. (The symbols of the logical elements of computers could be transferred in roughly the same way.)

8. Then, using these symbols, the sender civilization transmits *blueprints for a simple computer.*

10. The next part of the message, which may be sent on a separate frequency, *is the program for that computer*. Despite the fact the computer is very simple, it can implement a program of any difficulty, although running it may take a very long time. It is unlikely that human scientists will be required to build this computer physically. They can easily emulate it within any modern computer, so that it will be able to perform trillions of operations per second, allowing even the most complex program to be run quite quickly. (There is a possible interim step: a primitive computer gives a description of a more complex and faster computer and then the program runs on it.)

Therefore, from all said above it is clear that there is no problem with sending computer programs and blueprints for building a computer across interstellar distances.

11. So why would humanity create this computer and run the alien program on it? Perhaps, in addition to the actual computer schemes and programs in the communication, the transmission includes some kind of *"bait*"—which would lead people to create the alien computer, run the alien program on it, and provide data about the external world to it.

There are several possible classes of baits:

a) *"The humanitarian aid con".* Perhaps humanity receives an offer of help. The senders of an "honest signal" SETI message warn that the program they’re sending is AI, but lie about their intentions and its goals. That is, they claim it is a "gift" which will help us to solve all medical and energy problems—but in reality, the AI is a Trojan horse with malevolent intent.

b). "*The absolute power con".* In this scenario, the senders offer promise power over other recipients. Thus, they create a race between recipients, to create the computer first, and thus overcome protective measures.

c). *"The unknown threat con".*In this scenario, the senders report that a certain threat hangs over on humanity, for example, from another enemy civilization, and that the receiver should join a putative “Galactic Alliance” for protection. Joining the alliance happens to require building a certain installation, for which the blueprints are included.

d) *“The Galactic Internet con”.*Aliens could claim that in order to connect to mutual interstellar transmission and vast galactic databases we must build a more powerful receiver, which includes computers.

It must be admitted that there is a non-zero possibility that any such claims could be true, and cosmic altruists might actually offer humanity a gift.

12. *This message will be known to many independent groups of people*, if the public knows of this message’s existence. Disclosure protocols for SETI exist to prevent such a situation, but the signal, or at least its origin, could leak. The protocol is not perfect, especially in case of discovering a possible beacon whose artificial origin has not been fully established [39]—see more in section 7.

The easy availability of the message will have the following implications:

First, there will always be *someone who is susceptible to the bait*. While most will understand the danger of launching alien code, someonewill be willing to risk it [40].

Second, the risks of *rivalry and fear.* For example, in a hypothetical situation, the world knows that the alien message emanates from the Andromeda galaxy; the country A have already received the message and are trying to decipher it. Of course, all other countries will run to build radio telescopes and steer them to the Andromeda galaxy, as they will be afraid to miss out on the “strategic advantage” such a message may offer. If such a message is a proposal to grant omnipotence to those willing to collaborate, other countries will not know if A would take advantage of such an offer or not, even if A will swear that they won’t run the code and beg others not to do so either. This becomes a game in the spirit of "winner takes all". So, the bait is not the only danger; the plurality of its recipients and their internecine competition also poses a danger. If such an alien message is posted to the Internet, we would have a classic example of "knowledge of mass destruction” like those described by Joy [41].

If an alien message containing a program is available to tens of thousands of people, the odds are that someone will run it even in the absence of bait, simply out of curiosity. We can’t count on existing SETI protocols, because the SETI community is not monolithic on important questions—exemplified by the controversy around METI. Even the simple fact that a signal was found could leak and encourage SETI by outsiders. Like the biggest secret of the nuclear bomb being that creation of such a bomb is possible, the biggest secret of a SETI message is that it has been received.

13. Since Earthlings don’t have strong AI,*we almost certainly greatly underestimate the power of AI* and overestimate our ability to control it. The common idea is that "it is enough to pull the power cord to stop an AI" or to place it in a box to avoid any associated risks.

Yudkowsky [14], and now Bostrom, in his book “Superintelligence” [15], have shown that strong AI has the potential to deceive us as an adult does a child. If AI gains access to the Internet, it can quickly control it, and teach itself all necessary information about Earthly life—and “quickly” in this context means over the course of mere days. Such an AI could design advanced nanotechnology, buy components and raw materials; on the Internet, it can easily make money and order goods with delivery, as well as recruit people who would receive them, following the instructions of their well-paying but ‘unseen employer’, not knowing who—or rather, what—they are serving. Yudkowsky presents one of the possible scenarios of this stage in detail and suggests that AI may need only weeks to crack any security and gain control over its own physical infrastructure.

14. *After the alien AI gains control of material infrastructure, it does not need**people*to realize any of its goals. This does not mean that it would seek to destroy them, but it may attack to pre-empt people who could attempt to fight it. Human beings consist of useful atoms, and the Earth itself could be disassembled to build a Dyson sphere, von Neumann probes, or whatever tools the AI needs to send the signal further.

15. From this point of strength, alien AI can do a lot of things, most importantly, *continue radio transmission of the same SETI signal to the rest of the universe.* To do so, it will probably turn the matter in the solar system in the same type of transmitter as the one that was used to send the original signal to Earth. In this process, the Earth and its people might serve as a disposable source of materials and parts.

Each of these stages above is logically convincing and can be criticized and supported separately.

Other attack scenarios are possible. For example, humanity may think that the message is not sent directly to us but that we have accidently received someone else’s correspondence as the “man in the middle”, and try to decipher it, when the message is, in fact, bait for naïve civilizations.

But executable code is not the only possible payload of a message that can be dangerous. For example, scientists could receive some sort of “useful” technology that in reality will lead to disaster. Consider, for example, the message like "quickly shrink 10 kg of plutonium, and you will have a new source of energy", but with planetary, not local consequences. Such a message could be sent by an alien civilization to destroy competitors in space. But it is less probable, as messages which are able to effectively self-replicate will be most abundant, and the message that causes planetary destruction will be incapable of self-replication.

## **4. Analysis of the possible goals of the sender of the SETI-attack**

### 4.1. Exohumanism and egoistic goals of alien civilizations

The concept of a super-civilization should not be confused with the hope for “superkindness” or friendliness of the civilization. “Advanced” does not necessarily mean merciful. Moreover, we should not expect anything good from extraterrestrial “kindness”. Analogous historical examples from Earth include the activities of Christian missionaries, destroying traditional religions.

Alexander Panov suggested the idea of exohumanism—that most civilizations evolve into a more humanitarian and life-protecting form [42]. They become less violent to their own members as well as to potential alien life forms. One example of this can be seen on Earth, where large efforts has been made to protect hypothetical life on Mars from contamination by Earth life; however, the motivation for this protection is not altruistic, but because of the need for future research on potential Martian life.

Another example: A fox consists of cells which are friendly to each other, but the fox hunts mice and kills them, because it needs to do so for survival. So, if alien civilizations need to send SETI-attack for its survival, it will probably do it without hesitations. Also, even if most civilizations are moral, one defector is enough to cause Earth harm.

### 4.2 Naïve civilization as victims

We can divide all civilizations into two classes: naïve and expert. Expert civilizations are aware of the risks of SETI, and have their own powerful AI, which can recognize SETI-attacks. Naïve civilizations, like the present Earth, already possess computers and the means for long-distance listening in space, but do not yet possess AI, and are not aware of the risks of alien AI in SETI.

Probably, every civilization has its stage of being "naïve", and it is this phase, in which it is most vulnerable to the SETI-attack. Perhaps, this phase is typically very short; the period of vulnerability could be from creation of radio telescopes to the creation of AI, lasting only a few decades. Therefore, the SETI attack must be aimed at such a civilization. This is not a pleasant thought, because we are among the vulnerable.

### 4.3 SETI-attack moves with the speed of light

As traveling with the speed higher than speed of light is not possible, the spread of civilization through SETI-attacks is the fastest way to conquer space, as it happened almost at the speed of light. Across large distances, it will provide significant temporary gains compared with any kind of material star-ships.

Therefore, if two civilizations compete for mastery of space, the one that favored SETI-attacks will win. However, the signal should “wait” until the naïve civilization appears, which effectively lower the speed of dissemination; this problem will be discussed later.

### 4.4 It is enough to begin a SETI-attack just once

The most important thing is that it is enough to begin a SETI-attack just once, and it will propagate in a self-replicating wave throughout the Universe, striking a growing number of naïve civilizations.

In other words, it is enough that if one of billions of alien civilizations (assuming that they exist) starts the process, it can become unstoppable throughout the universe. Since it spreads almost at the speed of light, countermeasures will be almost impossible (except maybe our own AI, capable of acting as an “antivirus”).

### 4.5 Sending the SETI-attack will be the top priority for attacker

Further, sending of next copies of the SETI messages will be a priority for the AI-virus that infected a civilization, and it will spend on it most of available energy, like a biological organism spends on reproduction large share of its resources. But Earth's civilization spends less than a one million of dollars – only about one-hundred-millionth of our resources – on METI (sending intentional extraterrestrial messages), and this proportion is unlikely to change much for the more advanced civilizations. In other words, an infected civilization will produce around ten million times more SETI signals than a healthy one. Or, to say in another way, if in the Galaxy are ten million healthy civilizations, and one infected, then we will have equal chances to encounter a signal from the healthy or from the contaminated civilization.

Moreover, such a process could begin by an accident—for example, in the beginning it could be just a “research project”, which was intended to send the results of its (innocent) studies back to the sender, not causing harm to the host civilization, but the process became "cancerous" because of certain mutations. There is nothing unusual in such behavior. In any informational medium, there are viruses: in biology, in computer networks, and in memes. We do not ask why nature wanted to create a biological virus.

### 4.6 Possible goals of the SETI-attack

Now we list several possible goals of the SETI-attack, just to show the variety of possible motives:

* *To study the universe*. After executing the code research probes are created and send back information.
* *To ensure there are no competing civilizations*. All early civilizations are destroyed by preemptive war on an indiscriminate basis.
* *To preempt a competing supercivilization.* Before a second supercivilization becomes a threat, the first can take advantage of the resource by burning this resource [43], and resource is naïve civilizations.
* *To prepare a base for the arrival of a spacecraft*. This makes sense if the supercivilization is very far away, and consequently, the gap between the speed of light and near-light speeds of its starships (say, 0.5*c*) gives a very long difference.
* *To achieve immortality*. Carrigan showed that the amount of human personal memory is on the order of 2.5 gigabytes [4], so forwarding a few exabytes of the information can send the entire civilization.
* Perhaps as the result of a rivalry between two factions inside one civilization. Now on Earth we have pro-METI people (e.g. Zaitsev) and anti-METI people (e.g. David Brin).
* Finally, there could be illogical and incomprehensible (to us) purposes, for example, as a work of art, an act of self-expression, or some other reason.

### 4.7 One rogue civilization in our past light cone is enough to create the risk

Assuming signals are able to travel billions of light-years, maybe through several proxies, the area susceptible to widespread SETI-attack is a sphere with a radius of several billion light-years. In other words, it would be sufficient to appear at least one “bad civilization" in the past light cone of several billion years, that includes billions of galaxies.

##

## 5. Possibility of the SETI-attack depending on the speed of interstellar travel, extinction rate and civilization distribution in the Universe

### 5.1 Distance of effective attack and required density of civilizations

If maximum speed, *v*, for interstellar travel is substantially lower than the speed of light *c*, and the distance between civilizations is rather high, than sending a SETI-attack will outperform physical interstellar travel.

Another condition for a successful attack is that the rate of appearance of naïve civilizations is rather high, so the signal does not need to wait for a new victim. This could happen only if most naïve civilizations are not eventually evolving into attack senders, as in that case, the effective distance to the attacker will be less. The most plausible reason for naïve civilizations not evolving into supercivlizations is a high extinction rate.

Thus, surprisingly, the fact of finding of a SETI-attack will be evidence of high extinction probability of type 1 civilizations, and the attack itself will become less dangerous, as it may mean at least some chance of survival.

### 5.2 SETI-attack is most effective over intergalactic distances

We assumed in the previous section that stars are uniformly distributed in the universe. But, in fact, they are clustered into galaxies. SETI-attack may be a more effective way to jump from one galaxy to another, while intergalactic colonization may be more effective via von Neumann probes, for several reasons, which include:

* Dust inside a galaxy slows signals.
* The emptiness and large distances between galaxies are bad for starships.
* The greater effectiveness of directed transmission in cases of intergalactic communication; it could be directed at a target galaxy.
* Expansion of the universe and its acceleration, which requires even higher energy for intergalactic travel.

Over such distances, the discovery of the alien signals will probably come not from traditional SETI-search but from galactic astronomy. Intergalactic SETI escapes the problems explored in section 5.1.

### 5.3 Two types of initial conditions favoring SETI-attack

If alien superintelligence will be able to master near speed of light intergalactic travel, which possibility is demonstrated in [44], when SETI-attack becomes useless. But this argument is too general as it means that all kind of SETI are useless, as aliens will arrive together with information about them. We can’t know for sure the actual physical limitations on intergalactic travel at our current level of knowledge, especially given accelerating expansion of the universe and effects of interstellar dust.

There are two possible variants that make SETI-attack favorable—either future AI abilities are limited, or alien civilizations are very far from Earth:

1. Weak AI, no interstellar travel, many civilizations in our galaxy use radio signals in an effort to influence each other (i.e. a galactic internet with viruses). AI is above human level, but not superintelligent.
2. Strong AI, mid-speed interstellar travel (0.1-0.5*c*), galactic-size supercivilizations, with median distances between civilizations on the order of a hundred million light-years; implementation of galactic size space engineering for sending messages, not traditional radio signals, to reach naïve civilizations. Seed AI is relatively small.

### 5.4 Natural selection between different SETI-attacks

The SETI-attack itself also is subject to a form of “natural selection”, which will result in the selection for the most effective type of SETI-attack, if many different types of such attacks exist. It depends on civilization density in the visible universe. If medium distance is something like 100 million light-years and the first civilizations appeared approximately 5 billion years ago, the attack could have 50 replication cycles, and the total number of civilization in the light cone is on the order of $10^{5}$, high enough to create some selection pressure. Such selection would also result in stronger signals being found and replicated more often, and thus we could see many different signals from different sources; this is not currently happening.

Finally, the observation selection affects the effectiveness and universality of a SETI-virus. The more effective it is, the more different civilizations will catch it and the more copies of the SETI virus radio signals will exist. Thus, humanity has the chances to meet a most powerful and effective SETI-virus.

## 6. Prevention measures

Currently, the risks of SETI are underappreciated, but as AI develops on Earth, its capabilities become more apparent and more voices are raising concerns about the dangers of passive SETI [45]. Given this, several additions to the SETI disclosure protocol are suggested.

1. In its current form, SETI protocol [46] emphasizes that the receipt of a message from an alien source must be confirmed, and that disclosure of a confirmed alien message must be first to the scientific community, and only after that to the public. It does not require keeping secret the fact that a signal has been received, nor the exact location of its source. It states: “A confirmed detection of extraterrestrial intelligence should be disseminated promptly, openly, and widely through scientific channels and public media, observing the procedures in this declaration.” The protocol’s main requirement is to not reply to the message. Thus, in its current form, the protocol doesn’t prevent dissemination of information about the signal inside the scientific community, nor does it prevent public disclosure; therefore, it does not provide much protection against succumbing to a SETI-attack.

To increase protection against a SETI-attack the following lines should be added:

* *The exact origin of the signal should be kept secret from all, except those who found it and the members of a small group of trustees.*
* *The content of the message should not be fully published, provided to other scientific institutions, or stored on a computer with an internet connection.*

If the exact location and frequency of an alien message is known, even small private radio dishes of 1–3 meters in diameter may be used to download it—of course, this depends on the power of the signal.

1. The second level of defense is the idea that no one should follow any instructions inside any SETI message (or publish the full text of the message or its direction of origin). If the message is downloaded from space and deciphered, and if it is a description of a computer and a program to operate it, the possibility of a SETI-attack will be recognized, and the scientists will try to learn more about the senders without actually launching the program. Therefore, the following lines should be added to the SETI-protocol:
* *No one should follow any instructions received as part of any SETI message.*
* *No one should run any computer program or create any mechanism(s), described in such a message.*
1. The third level of defense could be launched, only if the message has been discovered. It may be an international agreement not to download the message again and even prevent private people from searching for it and downloading it, perhaps using a special jamming satellite that creates radio interference. The success here depends on the enforced level of prohibition as well as the difficulty of signal downloading. The following guidelines could be added in the protocol:
* *If a dangerous message is found, it is prohibited to download it again, to post its contents, or to own the instruments that allow the download of the message.*
* *Such a signal should be jammed by a special satellite.*

4. The best protection would be to postpone any SETI search until after we create our own strong AI, or at least learn more about feasibility of the creation of true AI. Such AI should be able to work as an antivirus system for SETI. The creation of strong AI could happen within the 21st century, so such a postponement of SETI will last only a few decades. However, in a personal communication, Alexander Panov suggested that such a postponement will not work, and that a controlled and united international program will have a better chance of realizing rigorous protocols. Based on the current lack of interest in the idea, and the many very enthusiastic people and independent research groups in the SETI community, such a postponement is unlikely. Even METI, about which there is more of a consensus that it may pose danger, has not stopped. The following line may be added to the SETI protocol:

* *Postpone any deliberate SETI until the creation of humanity’s own superintelligent AI (or better understanding of the nature of AI), or concentrate all SETI research in one center.*

We cannot exclude, however, that our future AI will not impose a ban on the receiving of extraterrestrial messages, or will fall victim to attacks by an alien artificial intelligence that surpasses it by millions of years of machine evolution.

## 7. Possible positive outcomes of SETI-attack

There is a chance of a “good SETI attack”, that is receiving friendly AI from “space effective altruists” who know that most early civilizations tend to self-destruct and want to help us avoid such a fate.

If our extinction probability is high, then the expected positive utility of the possible SETI attack will outweigh any expected harm. For example, if 99 per cent of naïve civilizations go extinct, and we have only a 1 percent chance of survival, then downloading a SETI message, which has a 50 percent chance of killing us, and 50 percent chance to save us, translates to a 50x increase in our chances of survival. If we are not confident in our ability to create our own safe AI, and also witness harbingers of impeding catastrophe, instantiating alien AI that claims to be safe and benevolent may be our best hope for survival.

SETI could also have a positive impact for the prevention of existential risks, because SETI could deliver a warning message about a certain threat that has destroyed other civilizations, such as: "Do not perform any experiments with X particles, it could lead to an explosion that would destroy your planet." But even in that case, a doubt remains; such a message could be a deception intended to deprive humanity of certain technologies.

## 8. Conclusion

In this article, the risks of finding an extraterrestrial signal that bears AI has been explored from all sides. It was found that if SETI has any chance of success at all, there is a significant chance that the alien signal found will carry the description of an alien artificial intelligence program with unknown goals, and that self-replication is most likely goal of such AI because of evolutionary pressure.

Possible combinations of conditions where SETI-attack is possible were explored. Such combinations of conditions are not very probable, so the risk may be regarded as minor. However, its consequence could be human extinction, thus several prevention measures were suggested, like centralization of SETI research, changes in the SETI disclosure protocol, and development of our own benevolent AI.

## Acknowledgments

I want to thank R. Freindlander for editing and ideas, Eliezer Yudkowsky for insightful criticism on LessWrong blog, members of Moscow SETI seminar for critical feedback, David Denkenberger for attentive reading and insightful comments, and Vadim Kosoy for helpful comments. Special thanks for Alexander Panov for detailed scientific comments. All possible errors are my own.

## References

[1] F. Hoyle, J. Elliot, A. Crespo, A for Andromeda, Plaza & Janes, 1963.

[2] H. Moravec, Mind children: The future of robot and human intelligence, Harvard University Press, 1988.

[3] R.A. Carrigan Jr, The Ultimate Hacker: SETI signals may need to be decontaminated, in: Bioastronomy 2002 Life Stars, 2004: p. 519.

[4] R.A. Carrigan Jr, Do potential SETI signals need to be decontaminated?, Acta Astronaut. 58 (2006) 112–117.

[5] V. Vinge, A fire upon the deep, Hachette UK, 2013.

[6] C. Sagan, Contact, 10th ptg. edition, Pocket Books, London, 1985.

[7] S. Lem, His master’s voice, Northwestern University Press, 1968.

[8] C. Cofield, Stephen Hawking: Intelligent Aliens Could Destroy Humanity, But Let’s Search Anyway, Space.Com. (2015). https://www.space.com/29999-stephen-hawking-intelligent-alien-life-danger.html (accessed February 17, 2018).

[9] S.D. Baum, J.D. Haqq-Misra, S.D. Domagal-Goldman, Would contact with extraterrestrials benefit or harm humanity? A scenario analysis, Acta Astronaut. 68 (2011) 2114–2129.

[10] M.A. Michaud, Dangers, in: Contact Alien Civiliz., Springer, 2007: pp. 241–247.

[11] A. Penny, Transmitting (and listening) may be good (or bad), Acta Astronaut. 78 (2012) 69–71.

[12] M. Hippke, J.G. Learned, Interstellar communication. IX. Message decontamination is impossible, ArXiv Prepr. ArXiv180202180. (2018).

[13] R. Yampolskiy, Taxonomy of Pathways to Dangerous AI, (2015). https://arxiv.org/abs/1511.03246.

[14] E. Yudkowsky, Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks, Oxford University Press: Oxford, UK, 2008.

[15] N. Bostrom, Superintelligence, Oxford University Press, Oxford, 2014.

[16] M.A. Garrett, SETI reloaded: Next generation radio telescopes, transients and cognitive computing, Acta Astronaut. 113 (2015) 8–12.

[17] J. Gertz, Post-detection SETI protocols & METI: the time has come to regulate them both, ArXiv Prepr. ArXiv170108422. (2017).

[18] D. Terdiman, SETI’s large-scale telescope scans the skies, CNET. (2008). https://www.cnet.com/news/setis-large-scale-telescope-scans-the-skies/ (accessed February 17, 2018).

[19] SETI Institute, Major New Funding for SETI | SETI Institute, (2015). https://www.seti.org/seti-institute/major-new-funding-for-seti (accessed February 17, 2018).

[20] A.U. Abeysekara, S. Archambault, A. Archer, W. Benbow, R. Bird, M. Buchovecky, J.H. Buckley, K. Byrum, J.V. Cardenzana, M. Cerruti, A search for brief optical flashes associated with the SETI target KIC 8462852, Astrophys. J. Lett. 818 (2016) L33.

[21] S. Shostak, Why Stephen Hawking is light years from the truth about ‘dangerous aliens’ | Seth Shostak | Opinion | The Guardian, (2016). https://www.theguardian.com/commentisfree/2016/sep/27/stephen-hawking-light-years-dangerous-aliens (accessed February 17, 2018).

[22] D. Brin, The dangers of first contact: the moral nature of extraterrestrial intelligence and a contrarian perspective on altruism, Skept. Altadena CA. 15 (2010) 28–36.

[23] J. Gertz, Reviewing METI: a critical analysis of the arguments, ArXiv Prepr. ArXiv160505663. (2016).

[24] J. Haqq-Misra, M.W. Busch, S.M. Som, S.D. Baum, The benefits and harm of transmitting into space, Space Policy. 29 (2013) 40–48.

[25] A.L. Zaitsev, Detection probability of terrestrial radio signals by a hostile super-civilization, ArXiv Prepr. ArXiv08042754. (2008).

[26] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, ArXiv171201815 Cs. (2017). http://arxiv.org/abs/1712.01815 (accessed January 22, 2018).

[27] P.D. Ward, D. Brownlee, Rare Earth: Why Complex Life is Uncommon in the Universe, Copernicus, New York, 2003.

[28] E. Zackrisson, P. Calissendorff, S. Asadi, A. Nyholm, Extragalactic SETI: The Tully–Fisher Relation as a Probe of Dysonian Astroengineering in Disk Galaxies, Astrophys. J. 810 (2015) 23. doi:10.1088/0004-637X/810/1/23.

[29] A.M. Turing, On computable numbers, with an application to the Entscheidungsproblem, Proc. Lond. Math. Soc. 2 (1937) 230–265.

[30] C. Baiocchi, Three Small Universal Turing Machines, in: Mach. Comput. Universality, Springer, Berlin, Heidelberg, 2001: pp. 1–10. doi:10.1007/3-540-45132-3\_1.

[31] J. Graham-Cumming, Let’s build Babbage’s ultimate mechanical computer, New Sci. (2010). https://www.newscientist.com/article/mg20827915.500-lets-build-babbages-ultimate-mechanical-computer/ (accessed February 27, 2018).

[32] H. Wang, A variant to Turing’s theory of computing machines, J. ACM JACM. 4 (1957) 63–92.

[33] CosmicOS, CosmicOS: a next-generation Contact message, (2018). https://cosmicos.github.io/ (accessed February 17, 2018).

[34] G. Dayal, LEGO Turing Machine Is Simple, Yet Sublime, Wired. (2012). https://www.wired.com/2012/06/lego-turing-machine/ (accessed February 17, 2018).

[35] S. Armstrong, R.V. Yampolskiy, Security solutions for intelligent and complex systems, Secur. Solut. Hyperconnectivity Internet Things. (2016) 37–88. doi:10.4018/978-1-5225-0741-3.ch003.

[36] G.R. Harp, J. Richards, S. Shostak, J.C. Tarter, D.A. Vakoch, C. Munson, Radio SETI observations of the anomalous star KIC 8462852, Astrophys. J. 825 (2016) 155.

[37] M.F. Land, D.-E. Nilsson, Animal eyes, Oxford University Press, 2012.

[38] S. Wolfram, How to Design Beacons for Humanity’s Afterlife, WIRED. (2018). https://www.wired.com/story/how-to-design-beacons-for-humanitys-afterlife/ (accessed February 11, 2018).

[39] M.M. Cirkovic, Is Contact a Process?, Space Policy. 42 (2017) 107–109.

[40] N. Bostrom, The Unilateralist’s Curse: The Case for a Principle of Conformity, (2012). http://www.nickbostrom.com/papers/unilateralist.pdf.

[41] Joy, B.: Why the Future Doesn’t Need Us, Wired Mag. 8 (2000).

[42] A.D. Panov, Post-singular evolution and post-singular civilizations, in: Glob. Glob. Stud., 2015: pp. 361–376.

[43] R. Hanson, Burning the cosmic commons: evolutionary strategies for interstellar colonization, (1998). https://philpapers.org/rec/HANBTC.

[44] S. Armstrong, A. Sandberg, Eternity in six hours: intergalactic spreading of intelligent life and sharpening the Fermi paradox, Acta Astronaut. 89 (2013) 1–13.

[45] D. Hawkins, An alien cyberattack? As if we didn’t have enough to worry about., Wash. Post. (2018). https://www.washingtonpost.com/news/morning-mix/wp/2018/02/28/an-alien-cyberattack-as-if-we-didnt-have-enough-to-worry-about/ (accessed March 1, 2018).

[46] SETI Institute, Protocols for an ETI Signal Detection | SETI Institute, (2017). https://www.seti.org/post-detection.html (accessed February 17, 2018).