Futures 101 (2018) 46–54

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

A Game of Stars: Active SETI, radical translation and the Hobbesian trap

Karim Jebari[a,*], Niklas Olsson-Yaouzis[b]

[a] Institute for Futures Studies, Holländargatan 13, 101 31 Stockholm, Sweden
[b] Department of Philosophy, Stockholm University, Stockholm, Sweden

ARTICLE INFO

Keywords:
Extra-terrestrial intelligence
Radical translation
SETI, existential risk
Game theory
Hobbesian trap
Stag hunt

ABSTRACT

Among scholars dedicated to *Search for Extra-terrestrial Intelligence* (SETI), the risks and possibilities of actively contacting extra-terrestrials (Active SETI) have been widely discussed. Yet, some fundamental philosophical problems concerning the possibility of translating an alien language have hardly been raised in this context. The proponents of Active SETI assume that, abswould an extra-terrestrial intelligent (ETI) entity choose to contact us, they would use radio signals to convey a coded message that would be possible for us to decode and translate. Furthermore, they argue, were we to transmit a message, then this message would also be possible to translate. However, any interstellar message would, for obvious reasons, be conveyed without context and without the possibility of meaningful interaction over timescales relevant to us.

According to the most influential research program in the philosophy of language, the meaning of an utterance is derived from its use in a context and is not intrinsic to the utterance by which it was conveyed. Therefore, while radical translation, i.e. learning an unknown language, is possible, it requires contextualized interaction. Only then can semantic behavior be observed, and utterances linked to meaning. Thus, merely an exchange of signals cannot produce meaningful communication. If this claim is true, there are important game-theoretical consequences of interstellar contact. An informal game theoretical analysis of this scenario, *A Game of Stars*, is described. This analysis suggests that the lack of communication may lead players into a *Hobbesian Trap*, where fear impels the players to a risk dominant strategy, potentially resulting in mutual destruction. Our conclusion is that interstellar contact is an underestimated existential risk. If true and given the relative ease of contacting an ETI given the knowledge of its location, information about the existence and location of an ETI would be very dangerous to spread. Thus, knowledge of an ETI and its location would constitute an information hazard.

1. Introduction

Although no evidence suggests that we will encounter extra-terrestrial intelligence (ETI) in the near future, it is a possibility that cannot yet be discarded (Shostak, 2004). We have no information as to how likely it is for life to emerge on a planet with the appropriate conditions, and no information as how likely or long-lived intelligent life is (Baum, Haqq-Misra, & Domagal-Goldman, 2011). According to the most "pessimistic" assumptions, we are the only civilization capable of interstellar communication in the Milky Way. According to "optimistic" assumptions regarding the probability of the emergence of life and sapience, we should expect

* Corresponding author.
E-mail addresses: karim.jebari@iffs.se (K. Jebari), niklas.olsson-yaouzis@philosophy.su.se (N. Olsson-Yaouzis).

https://doi.org/10.1016/j.futures.2018.06.007
Received 30 January 2018; Received in revised form 4 June 2018; Accepted 15 June 2018
Available online 18 June 2018
0016-3287/

more than a hundred million civilizations in our galaxy. Although this "optimistic" scenario is almost certainly false, it ought to focus our minds on the possibility of finding an ETI in our *stellar neighborhood*, here defined as a sphere centered on our planet with a radius of 100 light years. There are about 15 000 stars within 100 light years, assuming an average density of 0.14 stars/cubic parsec. Most of these stars are believed to be faint red dwarfs and are yet unknown. Current observations of extrasolar planets have given us reason to believe that earth-sized planets are much more common than previously thought, perhaps as common as one per five Sun-like stars (Petigura, Howard, & Marcy, 2013). The possibility of life on moons of gas giants (such as Jupiter's moon Europa, or Saturn's moon Enceladus) adds further reason to take the possibility of an ETI in the stellar neighborhood seriously.

Contact with an ETI may turn out to be among the most important events in the history of mankind. The worst possible outcome of this event is, as has been suggested by various scholars, an existential catastrophe, where human civilization meets the fate of the indigenous peoples of the Americas upon the arrival of Europeans (Baum et al., 2011). Therefore, such a possibility, even if unlikely, merits special scrutiny. We will present some considerations that have strategic relevance to a possible future interaction. This analysis will focus on a possible interaction with an ETI that is a type I civilization on the Kardashev scale. The Kardashev scale is a measure of a civilization's level of technological advancement based on how much energy it produces and consumes, where a type I civilization has an energy revenue equivalent to the insolation of an earthlike planet (Kardashev, 1964). Mankind is not yet a type I civilization but may become one over the next few centuries. A civilization at any higher level on the Kardashev scale (II-III) is no less likely to exist in our stellar neighborhood. However, would any such very advanced civilization in our neighborhood choose to attack us, there would be little that we could do about it. The game would be over before it started. We shall therefore focus on the potential risks associated with a type I civilization. This is not because such an encounter is more likely, but because the outcome of encounters with more advanced civilizations are not likely to be relevantly affected by human action.

First, we argue that translation between speakers with no common language (known as *radical* translation) requires repeated interaction in a familiar context. This claim is widely accepted among philosophers of language, although it seems to be overlooked among proponents of Active SETI. This implies, we argue, that translating a message conveyed to us by an encoded signal is not only very difficult, but *practically impossible*.

Second, we investigate the game theoretical implications of being unable to communicate with a type I ETI, with the technological prowess to cause us catastrophic harm. We argue that upon contact, mankind and the ETI risk being caught up in what game theorists call a *Hobbesian Trap*. This is a situation where uncertainty and fear direct the actions of the players, causing them to choose a risk dominant rather than a payoff dominant strategy in a game with more than one equilibria. In this game, the *Game of Stars*, a risk dominant equilibrium is a first strike, even when both players prefer peaceful co-existence.

Third, we defend this argument against the counterargument that we could infer that an ETI is unlikely to attack us from the mere fact that an ETI has the technological capability to make contact.

While we lack, as argued by Haqq-Misra (2018) certainty about the risks of contacting ETI, we argue that the risk profile of this choice is asymmetrical, with potential downsides that outweigh potential upsides. Thus, we conclude that actively contacting, as representatives from the SETI institute advocate (Vakoch, 2016), ETI or responding to an ETI signal would be *reckless*. Moreover, we argue that the knowledge of the existence and location of an ETI to the public should be considered as a major information hazard.

2. The decoding problem[1]

It has been argued that a Kardashev type I ETI (henceforth just ETI) that would be able to detect an encoded radio signal would also be able to decode it (Shostak, 2004). Moreover, NASA's SETI-program, where large arrays of radio telescopes await a signal from an ETI, is based on the premise that we could decode a radio signal sent to us from an ETI. For example, Devito and Oehrle argue that:

"Since it is likely that contact between our civilization and an alien one would be via radio, potential correspondents would have a basic knowledge of science. Such beings should therefore be able to learn a language based on fundamental science." (Devito & Oehrle, 1990).

We may refer to this claim as the "intelligibility claim": *Any entity that has the means to intercept a coded message from a distant planetary system will understand this message, without the need for additional context or interaction.*

This claim has been criticized in the literature for being too optimistic. For example (Atri, DeMarines, & Haqq-Misra, 2011) argue that most attempts at METI have been too anthropocentric and unlikely to be possible to decipher. Here, we wish to make a stronger claim. We will apply the theories of Willard van Orman Quine, one of the most influential philosophers of language of the 20th century. We argue that these theories imply that: *No entity can translate any message that humans could send with nothing but electromagnetic transmission.*

Quine presents a thought experiment known as *radical translation*. Here, a linguist encounters a native linguistic community whose language is completely unknown. Naturally, the linguist wants to learn this language and establish communication. How would the linguist proceed? Since the linguist cannot ask the native "what did you mean by X?", the linguist needs to use direct translation on occasion sentences. For example, after hearing the utterance "gavagai" many times as a rabbit is present, the linguist can hypothesize that "gavagai" means "rabbit", whereupon pointing at rabbits and non-rabbits and observing the behavior of the native can confirm the hypothesis. For example, by pointing at a potato, and hearing an utterance other than "gavagai", the linguist can reject the hypothesis that "gavagai" means food.

However, even when the linguist takes these first steps in creating a translation manual, there is still no certainty on whether

[1] This discussion draws on (Tennant, 2007).

"gavagai" means "rabbit". Gavagai could for example just as plausibly be translated as "undetached rabbit parts" or "rabbithood", since none of the observed evidence excludes these possible translations. To ascertain the correct translation of "gavagai", the linguist would need to understand the general structure of beliefs of the natives. However, such holistic understanding would require the correct translation of many words and concepts. The conclusion is that translation is indeterminate, which means that manuals for translating one language into another can be set up in many ways, all compatible with the totality of speech dispositions, yet incompatible with one another (Hylton, 2014; Quine, 1964).

Now it should be noted that this need not imply that communication must fail between the linguist and the natives. On the contrary, verbal behavior is often successful even when speakers cannot produce a complete translation manual. Yet, to create a rudimentary translation, any hypothesis of meaning can only be defended by appeals to assumptions about the psychology and social context of the speakers of the language and their specific context. For example, assume that the linguist has hypothesized that the correct translation of "gavagai" is "food." The linguist can test this hypothesis by uttering "gavagai" while pointing to objects that she believes that a native considers to be food, such as a potato. If the linguist believes that the native is cooperative, and if the native communicates what the linguist takes to be assent, then the linguist's hypothesis is corroborated; if the native communicates what the linguist takes to be dissent, then the linguist's hypothesis is falsified.

In Quine's thought experiment, the linguist interacts with a human, and is therefore familiar with many important aspects of human biology and psychology, for example that humans sometimes eat rabbits and potatoes. Such crucial information will not be available to a hypothetical astrolinguist. Moreover, to carry out radical translation our hypothetical astrolinguist would need to observe the speaker and how its utterances carry their meanings immediately (as in typical English utterances in contrast to coded signals) in a context (the speaker interacting with an environment). There is simply no inherent meaning in the utterance itself. This essentially behaviorist view of language remains prominent in contemporary philosophy of language:

Each of us learns his language by observing other people's verbal behavior and having his own faltering verbal behavior observed and reinforced or corrected by others. We depend strictly on observable behavior in observable situations. […] There is nothing in linguistic meaning beyond what is to be gleaned from overt behavior in observable circumstances (Quine, 1992, p. 37–38). While translation, or the exact and unambiguous analysis of meaning is an unattainable ideal, perhaps an *interpretation*, or "good enough translation" could suffice. The basic problem here is that one cannot assign meaning to utterances without knowing about the speaker's beliefs, while that knowledge is unattainable without understanding the utterances. To achieve such an interpretation, the philosopher Donald Davidson argues, our linguist needs to apply a *principle of charity* in the interpretation situation (Davidson, 1973). This implies that the linguist needs to ascribe as many true beliefs as possible to the alien speaker given the observed behavior and utterances and a general assumption of coherence in belief. In our thought experiment, the linguist can assume that the beliefs of the native are to a large extent in agreement with her own. Using those beliefs as guidance the linguist can gain a richer understanding of the beliefs of the speakers (Malpas, 2015).

However, in trying to interpret an ETI, an astrolinguist would lack important information about the alien speaker's perceptual apparatus and general belief structure. The astrolinguist in this case cannot even be certain that they are the intended recipient of the signal, or that it even is a message with semantic content, rather than a "ta-da!" that may or may not be intended to communicate "I'm here". Thus, the set of beliefs that can plausibly be ascribed to an ETI is arguably so minimal that it cannot be used to guide the process of interpretation.

Noam Chomsky used his theory of Universal Grammar (UG) to object to Quine's indeterminacy thesis (Chomsky, 1968). According to Chomsky, the fact that humans have an innate UG helps us pick out the correct translation manual among the set of compatible translation manuals. If our translations are guided by a reliable process (provided by UG), this would allow us to have knowledge about the correct translation (Pagin, 2000). However, we can't justify inter-species translation on the same grounds. Chomsky never claimed that UG was "universal" in the sense that it would be shared across all possible species able to use language. In other words, the linguist who meets human natives might be entitled to justify his or her translations by reference to a shared UG that allows the linguist to reliably pick out one translation among many. A linguist who meets a non-human, on the other hand, isn't (without additional arguments) entitled to justify his or her translations of the non-human language because it isn't clear that his or her UG reliably picks out the correct translation of the locutions in the source language.

While sentences in a language may be impossible to translate, perhaps images or even video could be an alternative (Vakoch, 2016). One idea is to transmit a large prime number N-squared as a series of 1 and 0 repeatedly. Then strings of N-squared signals for the on/off signals could fill in the pixels. This information could in theory be used to assemble a picture. However, pictures or even video do not escape the claim of underdetermination. Just as a sentence can plausibly be interpreted in a variety of ways, an image or an icon is just as prone to multiple inferences. Without knowledge of the ETIs language, culture, morphology or perceptual apparatus, no image could have a unique plausible translation. Consider the difficulty of archeologists to reach consensus with regards to the interpretations of ancient images in the absence of written materials. For example, only with the help of the Rosetta stone, a stone written in Egyptian hieroglyphs and in ancient Greek, were archeologists and linguists able to decipher and understand ancient Egyptian inscriptions. The fact that hieroglyphs were mostly figurative, representing stylized but recognizable objects did not make deciphering them easier. Neither did the fact that Egyptians were humans, that some aspects of their culture and history were known through other sources and that at least some of the corpus was accompanied by illustrated narratives telling us their history, make radical interpretation possible (Singh, 2000).

The difficulties in deciphering hieroglyphs can be instructive to the suggestion to use icons i.e. transmissions that are similar to what they seek to represent, as suggested by (Vakoch, 2008). He argues that information could be conveyed through transmissions of different wavelengths whose relative position can convey an idea or a concept. For example, at time t1, a transmission string is sent in one frequency and another string at a different frequency. Each string transmits a specific pattern with some intervals. At t2 a third

string is transmitted in an intermediate frequency with a combination of the patterns from the two strings. According to Vakoch, this is an iconic representation of sexual reproduction (Vakoch, 2008). However, such representations will suffer from the same problem that concerned Egyptologists. Similarity is not intrinsic in a way that allows for the inference suggested by Vakoch.

Among mathematicians and linguists that study the potential to communicate with ETI, it is often claimed that since a certain degree of mathematical understanding is required to construct radio emitters and receivers, ETI would share at least some of our fundamental understanding of the physical world (Devito & Oehrle, 1990). This common understanding would be common knowledge and form the basis of any communication attempts. This idea was for example the basis of Freudenthal's famous attempt at creating a self-explaining message (Freudenthal, 1960).

However, a common understanding of the world is not sufficient to establish radical translation. First, as Quine argued, theory is underdetermined by observation. This means that ETI could make all observations made by us, construct all technological devices produced by us, and still have a radically different theory of the world. Empirical observations typically force us to change our beliefs, but they don't tell us which of our beliefs that should be revised. (Quine, 1951). For example, Newton's theories of motion and acceleration are just as useful as General relativity for predicting the behavior of small objects moving at non-relativistic speeds. However, it is conceivable for a civilization to figure out relativity without having formulated a Newtonian theory of motion. In that case, both 19th century humans and this hypothetical civilization would have two distinct theories that would be equally good at predicting the behavior of small and slow-moving objects (Stanford, 2016).

Second, even if ETI had the very same scientific theories of the world, they would not necessarily share other non-scientific beliefs, such as "waterfalls are beautiful" and "ripe tomatoes should be enjoyed with olive oil". While science is undeniably an important aspect of human culture, scientific propositions such as "a hydrogen atom consists of a proton and an electron" form only a minor subset of an average person's (even an average scientist's) belief network.

Third, we have good reason to believe that the more technologically advanced a culture is, the more is it likely to differ from other cultures of the same technological level, other things being equal. Technology is a means for a species to gain control over the world. This means that the environmental and/or biological constraints that force relatively "primitive" civilizations to adopt similar adaptive strategies (terrace agriculture, the calendar) to solve practical problems are reduced. Two very advanced civilizations that develop independently will likely have solutions to the same problems that are to a greater degree dictated by founder effects, phylogenetic inertia and random drift than by the evolutionary forces imposed by nature. Thus, even if ETI and humans would have a similar biology and similar climactic conditions, we should expect an advanced ETI to be very alien to us and to have a correspondingly alien way of life. These differences would be radically amplified if the ETI would have a different biology and/or evolutionary environment.

To sum up, it is highly unlikely that our scientific theories are like a similarly advanced ETI's. And even if we would share some basic scientific understanding, it is highly unlikely that we would represent that shared understanding in the same way.

3. A game of stars

If we were to contact an ETI, by responding to one of their signals with a signal of our own, we will face a serious practical problem. Should we treat the ETI as a potential ally or as a potential enemy? Should we, so to speak, lower our defenses and welcome them with arms outstretched, or should we arm ourselves and strike before they get the chance to attack us? In this section, we will argue that assuming the practical impossibility to communicate with an ETI across interstellar distances will have severe implications for how we answer these questions. The nature of an encounter with ETI has been discussed previously, but much of this speculation has concerned whether ETI are more likely to be benevolent (Sagan & Newman, 1983) or malevolent (Ferriss, 2011). Seth Baum describes a set of scenarios where ETI endorses a form of universalist ethics (Baum, 2010). Our argument will not depend on any normative theory but will explore a human-ETI encounter in a game-theoretical framework, that is valid under the assumption that both players assign higher value to their own survival than to the survival of the other.

Consider first that any civilization that can send signals between the stars is probably able to cause great harm at a relatively small cost.[2] This means that we have reason to fear the weapons of the ETI, and that we have reasons to believe that the ETI fears our weapons. Next, note that even if we have no reason to believe that a newly encountered ETI is hostile, we don't have any immediate reasons to rule out that they are hostile; vice versa, we have no immediate reason to believe that they believe that we are not hostile.

This situation resembles the relationship between the US and Soviet during the cold war (Korhonen, 2013). The US and the Soviets had the ability to cause catastrophic harm while being unable to prevent such harm. Furthermore, the Americans didn't trust the Soviets to refrain from hostile actions, and they also believed that the Soviets didn't trust the Americans to refrain from hostile actions. Let us begin by using a game theoretical model to describe the problem faced by the nuclear super powers during the cold war and see what it can teach us about a hypothetical encounter with an ETI.

The relationship between the US and Soviet is often analyzed with game theory, the study of strategic interaction between *rational* players. A player is rational if, and only if, she has *consistent* beliefs, a *complete* and *transitive* preference ordering, and makes the play that will maximize her expected preference satisfaction. This means that the game-theoretic versions of the leaders of the US do not

[2] Consider the following possible weapon of mass destruction: A mass driver can, in the hard vacuum of space or the Moon surface, accelerate an object to relativistic speeds. A missile traveling at such speeds does not need a warhead. The payload of a Saturn V rocket (45 000 kg) would, if accelerated to 0.1 C, yield almost 5 gigatons of TNT in kinetic energy. At 0.5 C, the yield would be 150 gigatons. By comparison, the *Tsar Bomba*, the most powerful weapon ever built, had a yield of 0.05 gigatons. See also (Brin, 2014) for ideas of how such weapons may look like.

hold contradictory beliefs, such as, if we attack first we will win *and* if we attack first we will not win.[3] That their preference ordering is complete means that for each pair of available options, *x* and *y*, they will either prefer *x* to *y*, *y* to *x*, or be indifferent between *x* and *y*. That is, it is not the case that the American leaders find two options, such as world peace and communist domination incommensurable. That their preference ordering is transitive means that for any three options, *x*, *y*, and *z,* if *x* is preferred to *y* and *y* is preferred to *z*, then *x* is preferred to *z*. So, for example, if the leaders of the US prefer peaceful coexistence to successful pre-emptive strike, and successful pre-emptive strike to all-out war, then the leaders prefer peaceful coexistence to all-out war. A rational player will never choose a strategy, *s*, if she believes that *s* will lead to the outcome, *y*, and that there is an alternative strategy, *s'*, that will lead to the outcome, *x* that she prefers to *y*.

In game theory, it is also common to assume that the players have correct beliefs about the game they are playing. The leaders of both the US and Soviet, for example, are assumed to both correctly believe what will happen if they decide to attack the other's capital. Furthermore, it is often assumed that the player's rationality and the structure of the game is *common knowledge*. A proposition, *p*, is common knowledge between two players, *the US* and *the Soviets*, if both know that *p*, both know that both know *p*, both know that both know that both know that *p,* and so on *ad infinitum*.

These assumptions have been challenged because they do not provide an accurate description of real human agents. However, even if people are not usually rational in the sense game theory assumes, it can be argued that the agents involved in the cold war come close enough for the model to be useful. After all, since the leaders and advisors during the cold war meticulously analyzed and scrutinized each other's behavior, it can be assumed that they were rational in the sense game theory assumes.

That the players are rational does not mean that they will always do what would normally be considered "rational." To see this, assume that both the leaders of the US and Soviet prefer peace to war but that both, if there were to be war, prefer to strike before the other player, and that both prefer to be engaged in bilateral war over becoming the victim of a surprise attack. If it is assumed that the value of peace is 1, launching a surprise attack 0.5, bilateral war is 0, and becoming the victim of a surprise attack -0.5, their predicament can be represented by the following (standard form) game:

Game 1: Cold war		US	
		Attack	Don't attack
Soviet	Attack	**0, 0** (war)	0.5, -0.5 (Soviet surprise attack)
	Don't attack	−0.5, 0.5 (US surprise attack)	**1, 1** (peace)

Since both the US and Soviet prefer peace to war it might seem rational not to attack. Indeed, if the US does not attack, then the Soviet's best response is to not attack: under the assumption that the US does not attack, Soviet gets 1 if it does not attack and 0.5 if it attacks. However, note that if the US were to attack, then Soviet's best response is to attack as well: under the assumption that the US attacks, Soviet gets 0 if it attacks and -0.5 if it does not attack. Because the game is symmetric the same is true of the US. Thus, there are two pairs of strategies that are best responses to each other: (don't attack, don't attack) and (attack, attack). Pairs of strategies that are best response to each other are called Nash equilibrium.

If the game has a unique Nash equilibrium game theory predicts that two fully rational players will play their respective equilibrium strategy. However, if there is more than one Nash equilibrium it is not clear what rational agents will do. Game theorists distinguish between a *payoff dominant equilibrium* and a *risk dominant equilibrium* (Harsanyi, 1995). If there are two Nash equilibriums, **s1** and **s2**, and both players are better off in **s1** than in **s2**, then **s1** is the payoff dominant equilibrium.[4] In Game 1, "peace" is the payoff dominant equilibrium because both players are better off in this equilibrium. If a player trusts the other player to opt for this equilibrium it would indeed be irrational for that player not to play its payoff dominant strategy.

An equilibrium is risk dominant, on the other hand, if it is the "safer" of the two equilibriums. That is, if there is a great payoff-difference for a player P1 between playing her payoff dominant strategy and another strategy when the other player P2 does *not* play his payoff dominant strategy, then it is much "safer" for P1 to play the non-payoff dominant strategy. In other words, by playing a "safe" strategy P1 can ensure that the worst outcome (-0.5) does not obtain. If a player is uncertain about what the other player will do, then it may make sense to play the "safe" strategy even though it would not maximize the player's payoff. So, if there are two Nash equilibriums, **s1** and **s2**, and **s2** consists of two "safe" strategies, then **s2** is the *risk-dominant equilibrium*.[5] In Game 1 "war" is the risk dominant equilibrium. If Soviet does not attack it risks ending up with -0.5, whereas if it strikes first it ensures its payoff becomes at least 0. The more uncertain a player is about the other player's action the more likely she will be to play her risk dominant strategy.

By describing the equilibriums in terms of payoff and risk dominance it becomes clear that the question of where rational players will end up becomes a question of whether they trust each other to play the payoff dominant strategy. If they trust each other, it is likely that they will end up in the payoff dominant equilibrium. If not, it is likely that they will end up in the risk dominant

[3] To be more specific it means that they cannot be *Dutch booked*.

[4] Formally, assume that **s1** = (s1, s2) and **s2** = (s1',s2') are two Nash equilibrium. **s1** payoff dominates **s2** if, and only if, u1(**s1**) = > u1(**s2**) and u2(**s1**) = > u2(**s2**) and at least one of the inequalities are strict [u1(**s1**) > u1(**s2**) or u2(**s1**) > u2(**s2**)].

[5] In a symmetric two-player game with two Nash equilibriums, **s1** = (s1, s2) and **s2** = (s1',s2'), **s2** risk dominates **s1** if the expected payoff of playing s1' is at least as great as the expected payoff of playing s1 under the assumption that the other player flips a coin between s2' and s2.

equilibrium. These considerations led some game theorists to describe the problem faced by the superpowers as a *Hobbesian Trap*. Although both Soviet and US prefer peace, the lack of trust may force them to play the risk dominant strategy.

Moreover, once a seed of doubt is planted, a situation will quickly spiral out of control. Thomas Schelling describes this phenomenon in *The Strategy of Conflict* where he asks us to imagine that he, with a gun in his hand, discovers an armed burglar in his home. Should Schelling shoot or not? Here is Schelling's reasoning:

"Even if he [the burglar] prefers to just leave quietly, and I wish him to, there is a danger that he may *think* I want to shoot, and shoot first. Worse, there is danger that he may think that I think *he* wants to shoot. Or he may think that *I* think *he* thinks I want to shoot. And so on." (Schelling, 1981)

Applied to the standoff between the US and Soviet this means that the US may have to strike first if it suspects that Soviet is about to attack. This suspicion may be triggered by the belief that the Soviets think that the US is about to attack, or the belief that the Soviets think that the US thinks that the Soviets is about to attack, or that… In other words, as soon as uncertainty creeps into the strategic reasoning there is a real risk that the players end up in the risk dominant equilibrium, all-out war.

Arguably, it was this type of reasoning that culminated in the Cuban Missile Crisis and led the world to the brink of nuclear war in 1962. One of the lessons learned from this crisis was that direct communication between the superpowers' leaders prevents uncertainty from spiraling out of control. To facilitate direct communication between the superpowers the "Moscow-Washington hotline" was established.

Game theorists have also tested the intuitively plausible claim that communication helps players coordinate on the payoff dominant strategy in *Cold War*-like games. When played in a laboratory, the likelihood that the players successfully coordinate on the payoff dominant equilibrium increases when one of the players can announce her intended play. And if two-way communication is allowed, so that both players simultaneously announce what they intend to play, the likelihood of ending up in the payoff dominant equilibrium increases even further.[6]

The intuitive explanation of why communication helps in a cooperative game with multiple equilibriums is that it allows players to turn one of the equilibriums into a *focal point* (Schelling, 1981). By making the payoff dominant equilibrium a focal point player *A* becomes confident that the other player, *B*, is also focused on this equilibrium; *A* also becomes fairly confident that *B* believes that *A's* attention is focused on this equilibrium; *A* also becomes fairly confident that *B* believes that *A* believes that *B*'s attention is focused on this equilibrium; and so on. It should be noted that deterrence is an act that requires the ability to communicate. Merely showing one's ability to harm could be interpreted as an actual attempt at harming. Without having established communication, such attempts at deterrence are very risky and could cause the conflict they sought to prevent.

For our purposes, the upshot of this section is pessimistic. Given the conclusion that we cannot communicate with an ETI we might be forced to accept the following pessimistic argument:

1 If a player cannot trust the other player she should play her risk-dominant strategy.
2 Communication is necessary to build trust.
3 We cannot communicate with ETIs.
4 Therefore, we cannot build trust with an ETI.
5 Therefore, we should play our risk-dominant strategy.
6 Our risk dominant strategy is to attack first.
7 Therefore, we should attack first.

4. Signals and evolutionary paths

The conclusion that we ought to strike first might be premature. Even if we are unable to translate the ETI's signal we may be able to draw some conclusions about their intentions based on the mere fact that they have broadcasted an interstellar signal. That is, even if we don't understand what the signal means, we can use it as evidence to support other conclusions. The difference between what a signal *means* and what it *indicates* can be illustrated with a simple example. Imagine that you are marooned on what you believe is a deserted island. After a couple of days, you suddenly hear a cry, "jag är här!" If you know Swedish, you will understand that the cry *means* 'I'm here.' Even if you don't know Swedish and cannot figure out the meaning of the cry you will take it to indicate that someone else is on the island. Similarly, upon receiving an interstellar signal whose meaning we cannot discern, we are able to conclude that someone has sent an interstellar signal. On the deserted island, there is no reason to take the cry as evidence of peaceful intentions. After all, the cry may be an attempt to lure you into a trap. Similarly, the purpose of the interstellar signal may be to find civilizations to conquer. Thus, the mere fact that we have received an interstellar signal does not seem to provide conclusive reasons to refrain from attacking first.

However, it may be argued that there is a crucial difference between a person who cries "jag är här!" and someone who sends an interstellar signal. No skill (except rudimentary knowledge of Swedish) is needed to cry, "jag är här!". Sending an interstellar signal, on the other hand, requires a lot of technical knowledge. If it is impossible for a civilization to develop interstellar communication

[6] See, for example (Cooper, DeJong, Forsythe, & Ross, 1992). They show that in a coordination game, with a similar pay off structure as our *Cold War* game, 98% of the players chose the risk-dominant strategy (corresponding to "attack" in *Cold War*) if no communication was allowed. If one-way communication was allowed then the equivalent of "I will not attack" was announced in 87 % of the cases, and the payoff-dominant equilibrium (corresponding to (don't attack, don't attack)) was played in 60 % of the games. Finally, when two-way communication was allowed 150 out of 165 pairs announced that they play the payoff-dominant strategy ("don' attack") and 94 % of these pairs ended up in the payoff dominant equilibrium (corresponding to don't attack, don't attack).

without following some norm of non-aggression, then the signal indicates that the ETI follows a norm of non-aggression. If this is true, then the mere fact that an ETI sends an interstellar signal provides evidence that the ETI doesn't intend to attack first. *If* this is true, then, an interstellar signal would be like a one-way signal announcing, "we don't intend to attack." Furthermore, because the ETI can be expected to realize this as well, if we were to respond to the signal with an interstellar signal of our own, it would be like announcing that we don't intend to attack. Thus, our back-and-forth signaling would have the same trust-building effect as two-way communication in the Cold War scenario above.

The crucial premise is that the act of transmitting an interstellar signal is evidence that the ETI is governed by a norm of non-aggression. There are several scholars that have argued that groups whose members value cooperation (Alexander & Skyrms, 1999; Skyrms, 2003), truthfulness (Williams, 2004), and non-aggression do better than groups whose members don't (Axelrod & Dawkins, 2006). As groups with different norms compete, cooperative and non-aggressive groups will prevail. In other words, norms of non-aggression are a strong evolutionary attractor that will overcome random variations in starting conditions. Thus, non-aggressive norms will in the long run, dominate any sufficiently advanced civilization.

Now, the argument goes, because a civilization that sends interstellar signals has considerable technological and industrial prowess, and has existed for a long time, it must at some point along its evolutionary path have developed a norm of non-aggression. Thus, the fact that the ETI can send interstellar signals provides conclusive evidence that it values non-aggression and we can therefore trust it to not attack first.

There are some problems with this argument. First, it is questionable whether an ETI's norm of non-aggression extends to us. It would extend to us if it were an *unconditional* norm in the sense that it was activated every time a member of the ETI civilization encountered another sentient being. However, it is unlikely that cultural evolution through natural selection will produce unconditional pro-social norms of this kind. To see this, let us shift focus from the selection of norms on the *group* level, to the selection of norms on the *individual* level.

Assume that we have a population of individuals who from time to time are paired to play a game with payoffs and strategies like the Cold War scenario above. Assume that everyone is an unconditional aggressor and will always strike first. In this population, every individual will have an expected utility of 0. Furthermore, if it is assumed that reproductive success is proportional to expected payoff, then this population cannot be invaded by a small group of unconditional non-aggressors, who will never strike first, since they will have an even lower expected payoff (close to -0.5) than the aggressors.

However, assume that some individuals learn to distinguish between non-aggressors and aggressors and adopt a strategy where they strike first against aggressors and abstain from attacking other non-aggressors. The members of this sub-group will do better than the unconditional aggressors. If they are paired with an aggressor they will get an expected payoff of 0, which is the same expected payoff as the aggressor gets, but if they get paired with a non-aggressor they will get an expected payoff of 1, which is higher than the expected payoff for the aggressor. Thus, on average the members of the new group, let's call them conditional non-aggressors, do better than unconditional aggressors.

Because it is assumed that reproductive success is proportional to expected payoff, conditional non-aggressors will on average have more offspring than the aggressors. This means that the proportion of conditional non-aggressors in the group will increase over time until they dominate the population. In other words, it's possible that the members of successful groups are governed by some norm of conditional non-aggression rather than a norm of unconditional non-aggression.

If the ETI is governed by a norm of conditional non-aggression, rather than a norm of unconditional non-aggression, then there is no guarantee that they will extend their conditional non-aggression to us. So, if it is possible to develop interstellar signaling technology while being governed by a norm of conditional non-aggression, then an interstellar signal cannot be conclusive evidence that the ETI will abstain from a first strike. Given what we know about humans' tendency to trust members of their own group, and distrust members of other groups, this should come as no surprise.[7]

The second problem with the argument is that it assumes that the only way of acquiring the technology of interstellar signaling is by following an evolutionary path that goes via the development of a norm of non-aggression. If this assumption is false, then even if a norm of unconditional non-aggression was required to *develop* interstellar signaling we could not use the fact that an ETI sent an interstellar signal as conclusive evidence that the ETI will not strike first. There are alternative ways of acquiring the technology that does not require a norm of non-aggression. The aggressive ETI could have received it as a *gift* from another civilization. For example, in Ian M. Banks' novel *The Hydrogen Sonata* we are invited to imagine that some highly advanced civilizations give technological gifts to less sophisticated and more aggressive civilizations. The aggressive ETI could also have acquired the technology through *conquest* of another more technologically advanced civilization. For example, the Mongols were nomads and had relatively limited technology in siege warfare. However, as they subjugated more technologically advanced populations, they could co-opt their technology to the detriment of their foes. Finally, it is also possible that the ETI is *created* by a civilization that has developed interstellar signaling. For example, an artificial intelligence could acquire the technologies of its creators without acquiring norms of non-aggression.

Once these possibilities are considered it is difficult to maintain that interstellar signaling provides conclusive evidence that the ETI will not attack us first. After all, the point of the interstellar signal may simply be to find someone to conquer. The worry is aggravated if we consider that the ETI may have similar doubts about us. It may decide to strike first if they doubt that we are governed by a norm of non-aggression and therefore suspect that we might attack first. They may also decide to strike first if they believe that we doubt that they are governed by a norm of non-aggression and suspect that they will attack first, and so on. Once this

[7] As for example David Hume (Treatise: 3.2.1) observes, "An Englishman in Italy is a friend: A European in China; and perhaps a man would be beloved as such, were we to meet him in the moon."

is taken into consideration we're back where we started. Uncertainty about the motives of an ETI cannot be eliminated unless communication is possible.

Uncertainty about ETI should not be equated with uncertainty with the risk of Active SETI. Some scholars have argued that there could be both potential risks and benefits to establishing contact with ETI (Haqq-Misra, Busch, Som, & Baum, 2013). Others have argued that there are considerable risks in *not* contacting ETI (Korbitz, 2014). Assuming the analytical framework of Haqq-Misra et al., and if our analysis is correct, then risk clearly dominates opportunity when contacting an ETI. This act would be analogous with for example not acting to curb the emission of greenhouse gases. In other words, Active SETI is a *reckless* act, or an act that has an asymmetrical risk profile (one choice is more risky than the other) *and* where risk dominates opportunity (Hansson, 2009).

5. Information hazards

An information hazard is knowledge that by its very existence poses a considerable existential risk (Bostrom, 2011). The paradigmatic example of an information hazard is the "Deplorable Word" in the fictional world created by C.S. Lewis in his novel *The Magician's Nephew*. Uttering this word, which is possible for anyone who knows it, ends all life. The characteristic of an information hazard is that upon learning some piece of information, the threshold to cause catastrophic harm is lowered. While very few individuals are prepared to cause such harm, the more people have this knowledge, the greater the chance that someone will use it, either with malevolent intent or by mistake.

If our analysis is correct, contacting an ETI is reckless. However, knowing the exact location of ETI in our stellar neighborhood would make it possible for many actors and organizations to establish contact. The equipment and know-how to send a signal to nearby stars are common enough that were the locations of ETI public knowledge, it would only be a matter of time before someone was reckless enough to unilaterally establish contact. Such unilateral actions are more likely to take place if each agent capable of taking the action is to make a personal judgement call on whether to act (Bostrom, Douglas, & Sandberg, 2016).

Thus, we suggest that we ought to abstain from sending out a message to the stars, but also that if any one of us would acquire knowledge about the location of ETI, that we keep that knowledge to ourselves to the greatest extent possible.

6. Other objections

It has been argued that if there is an ETI with the ability to cause mankind considerable harm in our interstellar neighborhood, they are likely already aware of our existence, or will become so in the next few decades as the expanding shell of radio and radar transmissions reach their telescopes (Shostak, 2013). This claim has been contested by (Gertz, 2016)) and (Billingham & Benford, 2014) and others. We grant that if Shostak's claims are true, the concern outlined here would be less urgent than otherwise. However, it should be noted here that a signal from us not only informs ETI where we are (which, if Shostak is correct, would make no difference), but also informs them that *we know* where they are. Knowing this, there is a considerable risk that this can be a stepping stone on the path towards Schelling's trap. Thus, even if an ETI would know about our existence, it would be reckless to send a signal.

7. Concluding remarks

To sum up, we have argued the following: (1) Communication attempts with EM transmissions, as proposed by METI proponents, with ETI will fail because there will not be sufficient context and/or possibility for interaction for either side to assess the meaning of the other side's communications. (2) In the absence of successful communication, we cannot trust ETI, and ETI cannot trust us. (3) In the absence of trust, and given the ability to harm each other, it is likely that contact will lead to hostile escalation. (4) Therefore, we should not announce our presence and location. (5) Furthermore, if some humans detect ETI, they should hide the discovery to prevent other humans from attempting to interact with the ETI.

Thucydides argued that there are three reasons for war: greed, fear and honor. As space is so vast, and space travel so expensive even for type I civilizations, greed is an unlikely cause for interstellar war, at least between type I civilizations. Yet fear could, unless checked by honor, cause war.

References

Alexander, J., & Skyrms, B. (1999). Bargaining with neighbors: Is justice contagious? *Journal of Philosophy, 96*, 588–598.

Atri, D., DeMarines, J., & Haqq-Misra, J. (2011). A protocol for messaging to extraterrestrial intelligence. *Space Policy, 27*, 165–169. http://dx.doi.org/10.1016/j.spacepol.2011.01.001.

Axelrod, R., & Dawkins, R. (2006). *The evolution of cooperation: Revised edition* (revised edition). New York: Basic Books.

Baum, S. D. (2010). Universalist ethics in extraterrestrial encounter. *Acta Astronautica, 66*, 617–623. http://dx.doi.org/10.1016/j.actaastro.2009.07.003.

Baum, S. D., Haqq-Misra, J. D., & Domagal-Goldman, S. D. (2011). Would contact with extraterrestrials benefit or harm humanity? A scenario analysis. *Acta Astronautica, 68*, 2114–2129. http://dx.doi.org/10.1016/j.actaastro.2010.10.012.

Billingham, J., & Benford, J. (2014). Costs and difficulties of interstellar "Messaging" and the need for International debate on potential risks. *Journal of the British Interplanetary Society, 67*, 17–23.

Bostrom, N. (2011). Information hazards: a typology of potential harms from knowledge. *Review Contemporary Philosophy,* 44–79.

Bostrom, N., Douglas, T., & Sandberg, A. (2016). The unilateralist's curse and the case for a principle of conformity. *Social Epistemology, 30*, 350–371. http://dx.doi.org/10.1080/02691728.2015.1108373.

Brin, D. (2014). The search for extraterrestrial intelligence (SETI) and whether to send "Messages" (METI): A case for conversation, patience and due diligence. *Journal of the British Interplanetary Society, 67*, 8–16.

Chomsky, N. (1968). Quine's empirical assumptions. *Synthese, 19*, 53–68.

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1992). Communication in coordination games. *Quarterly Journal of Economics, 107*, 739–771. http://dx.doi.org/10.2307/2118488.
Davidson, D. (1973). Radical interpretation. *Dialectica, 27*, 313–328. http://dx.doi.org/10.1111/j.1746-8361.1973.tb00623.x.
Devito, C. L., & Oehrle, R. T. (1990). A language based on the fundamental facts of science. *Journal of the British Interplanetary Society, 43*, 561–568.
Ferriss, T. (2011). *The mind's sky: Human intelligence in a cosmic context.* Random House.
Freudenthal, H. (1960). *North Holland Publishing Co. Lincos: Design of a language for cosmic intercourse*.
Gertz, J. (2016). *Reviewing METI: A critical analysis of the arguments.* ArXiv160505663 Astro-Ph Physicsphysics.
Hansson, S. O. (2009). From the casino to the jungle. *Synthese, 168*, 423–432. http://dx.doi.org/10.1007/s11229-008-9444-1.
Haqq-Misra, J. (2018). Policy options for the radio detectability of earth. *Futures*. http://dx.doi.org/10.1016/j.futures.2018.04.002.
Haqq-Misra, J., Busch, M. W., Som, S. M., & Baum, S. D. (2013). The benefits and harm of transmitting into space. *Space Policy, 29*, 40–48. http://dx.doi.org/10.1016/j.spacepol.2012.11.006.
Harsanyi, J. (1995). A new theory of equilibrium selection for games with complete information. *Games Economy Behaviour, 8*, 91–122.
Hylton, P. (2014). Willard van orman quine. In E. N. Zalta (Ed.). *The stanford encyclopedia of philosophy*.
Kardashev, N. S. (1964). Transmission of information by extraterrestrial civilizations. *Soviet Astronomy, 8*, 217.
Korbitz, A. (2014). *The precautionary principle: Egoism, altruism, and the active SETI debate. Extraterrestrial altruism, the frontiers collection*. Berlin, Heidelberg: Springer111–127. http://dx.doi.org/10.1007/978-3-642-37750-1_8.
Korhonen, J. M. (2013). MAD with aliens? Interstellar deterrence and its implications. *Acta Astronautica, 86*, 201–210. http://dx.doi.org/10.1016/j.actaastro.2013.01.016.
Malpas, J. (2015). Donald davidson. In E. N. Zalta (Ed.). *The stanford encyclopedia of philosophy*.
Pagin, P. (2000). Publicness and indeterminacy. In A. Orenstein, & P. Kotatko (Eds.). *Knowledge, language and logic: Questions for quine* (pp. 163–180). Kluwer Academic Print on Demand.
Petigura, E. A., Howard, A. W., & Marcy, G. W. (2013). Prevalence of earth-size planets orbiting sun-like stars. *Proceedings of the National Academy of Science, 110*, 19273–19278. http://dx.doi.org/10.1073/pnas.1319909110.
Quine, W. V. O. (1992). *Pursuit of truth: Revised edition* (revised edition). Cambridge, Mass: Harvard University Press.
Quine, W. V. O. (1964). *Word and object, studies in communication, first printing edition.* MIT Press.
Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophy Review, 60*, 20–43.
Sagan, C., & Newman, W. I. (1983). The solipsist approach to extraterrestrial intelligence. *Quarterly Journal Royal Astronomical Society, 24*, 113.
Schelling, T. C. (1981). *The strategy of conflict, reprint edition.* Cambridge, Mass: Harvard University Press.
Shostak, S. (2013). Are transmissions to space dangerous? *International Journal of Astrobiology, 12*, 17–20. http://dx.doi.org/10.1017/S1473550412000274.
Shostak, S. (2004). When will we detect the extraterrestrials? *Acta Astronautica, 55*, 753–758. http://dx.doi.org/10.1016/j.actaastro.2004.05.023.
Singh, S. (2000). *The code book: The science of secrecy from ancient Egypt to quantum cryptography* (Reprint Edition). Anchor.
Skyrms, B. (2003). *The stag hunt and the evolution of social structure.* Cambridge, UK; New York: Cambridge University Press.
Stanford, K. (2016). Underdetermination of scientific theory. In E. N. Zalta (Ed.). *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.
Tennant, N. (2007). The decoding problem: Do we need to search for extra terrestrial intelligence in order to search for extraterrestrial intelligence? In B. Gertler, & L. Shapiro (Eds.). *Arguing about the min.*. Routledge.
Vakoch, D. A. (2016). In defence of METI. *Nature Physics, 12*. http://dx.doi.org/10.1038/nphys3897 890–890.
Vakoch, D. A. (2008). Representing culture in interstellar messages. *Acta Astronautica, 63*, 657–664. http://dx.doi.org/10.1016/j.actaastro.2008.05.011.
Williams, B. (2004). *Truth and truthfulness: An essay in genealogy, 1.3.2004 edition.* Princeton, NJ: Princeton University Press.