Cooperative Signalling

Philosophy 444

21 October, 2019

Simple Signalling

Our basic signalling game was a 2-by-2-by-2 game.[1] It had

- 2 possible states of the world, one of which is revealed to sender.
- 2 possible messages that sender can transmit.
- 2 possible actions for receiver to take upon receipt.

Intuitively for now, we're interested in situations where one of the actions is best in one state of the world, and the other is best in the other state of the world. (And for now we're dealing with cooperative games, so it's 'best' for both players.)

It's worth thinking about how little capacity two things need to be in order to be able to play this game. Sender just needs the capacity to differentially respond to some feature of the world. It needs all the intelligence of a key on a keyboard. (Not the fancy circuitry the key is connected to - literally the key itself, which is pressure sensitive.) And receiver doesn't need much more than that. As long as it can reliably respond to two distinct signals, and those responses could if needed be distinct, you're good to play this game.

The last thing you need to get an evolutionary story going is that the things playing the game are subject to evolutionary pressures. That does rule out things like keys on keyboards, but it doesn't rule out very much in the natural world. Lots of things are subject to evolutionary pressures.

Let's add one last thing to the setup. The two states are more or less equally probable. That isn't always the case, but it will sometimes be the case at least.

Given that minimal setup, the philosopher of biology Brian Skyrms showed that, with probability almost 1, one of the separating equilibria for the signalling game will arise. That is, signalling will eventually happen. I don't know if this is actually the story of how signalling arose, but it seems plausible. (I certainly don't know a better story.)

The problem is that as soon as you get away from this very special case, then the mathematical results aren't so neat. If the two states are not equally probable, then plenty of plausible models end up converging to no signal being sent, and the 'receiver' acting as if the more probable state has obtained.

If it is a 3-by-3-by-3 game, then some plausible models converge to a 'partially pooling equilibrium'. Here is how Huttegger et al describe one such partial pooling equilibrium

> Consider a … signaling game with [three states], where the sender always sends signal 1 in both states 1 and 2, and who in state 3 sometimes sends signal 2 and sometimes sends signal 3. Pair this sender with a receiver, who does act 3 in response to both signals 2 and 3, and who upon receiving signal 1 sometimes does act 1 and sometimes act 2, as shown in Fig. 1. In this equilibrium, information about state 3 is transmitted perfectly, but states 1 and 2 are "pooled".

[1] A lot of these notes draws on two papers by Simon Huttegger and collaborations: Some dynamics of signaling games, and Evolutionary dynamics of Lewis signaling games: signaling systems vs. partial pooling.

What they show is that under some common models for how populations evolve, sometimes that is what the population evolves to. It isn't common (it was 4.7% of the time on one of their models), but it happens. But if you allow more mutations into the model, this tends to go away, and the pure signalling equilibrium becomes (yet) more likely to evolve.

But still, if the neat story becomes less neat with just the move from 2 states to 3 states, it becomes a little worrying what it's like for the messy real world.

Restricted Signals

Let's drop the assumption that there are as many messages as states. In particular, let's think about how to manage the 4-by-2-by-4 game. That's a game where there are

- 4 possible states of the world.
- 2 possible messages that can be sent.
- 4 possible actions to be taken.

Intuitively, what should we want to have happen here?

You might think at first the answer will be, "Assign one message to two of the states, and the other message to the other two, and then we'll be done." But it's a bit more complicated than that. Imagine, for example, that the states are equiprobable, and these are the payoffs. (I'll just list one, because it's a cooperative game still.)

	S1	S2	S3	S4
A1	3	2	1	0
A2	2	3	2	1
A3	1	2	3	2
A4	0	1	2	3

Then it is really important that you have one signal for S1/S2, and another signal for S3/S4. That will have an average payoff of 2.5. (Question for readers: Why?) And no other messaging system will have as high a payoff. For instance, if you have one signal for S2/S3 and another for S1/S4, then the average payoff will be just 1.75. (Again, it's worth thinking about why.)

So we want signalling systems where like states get similar signals, not ones where you use a common signal for S1 and S4. Happily, that is mostly what we see in real-world signalling systems.

But you don't always want to divide things up two ways. Imagine that this was the payoff table, and again you have just two possible signals.

	S1	S2	S3	S4
A1	8	0	0	0
A2	0	2	1	0
A3	0	1	2	1
A4	0	0	1	2

The optimal signalling strategy is to use one signal for S1, and the other for S2/S3/S4. Question: How should

hearer respond to these signals? Question: What's the expected payoff of these signals? We want our signals to mark practically salient differences in the world. Again, it is arguable that this is what we find.

In game theory we typically assume that everyone knows the underlying probability distribution, and the underlying payoff structure. In the real world, that's not always the case. Let's imagine that people don't exactly know the payoffs for various actions, and they don't exactly know the probability distribution over the states. But they do know that a speaker has a limited number of signals, and is choosing to send a signal that is optimised to their (i.e., the speaker's) beliefs about the probabilities and the payoffs. What will happen?

Well, arguably what will happen is that we'll get a signal that is somewhat **vague**. In the real world, when you hear someone described as 'tall', or 'rich', or 'smart', it is clear that they are being described as being towards the upper end of the height/wealth/intelligence spectrum. But how close to the top must they be for this description to be right? It seems that you can be a perfectly competent speaker of English and not really know. One recent hypothesis (developed most extensively by Cailin O'Connor at UC Irvine) is that vagueness arises because players in the signalling game don't know exactly the parameters of the game they are playing. And the optimal strategy, i.e., what states you pick out by your signal, is dependent on the precise values of these parameters. There has been a lot of work in philosophy and linguistics about what vague terms mean, but a lot of it treats vagueness as some kind of defect of the language, as if it's weird why it is even there. This is a very interesting proposal for why we would naturally have ended up with vague language.

Lewis on the Development of Language

A lot of the work on this topic nowadays is done by economists and, especially, biologists. But it turns out that the origin of the work is in the early work by the most influential Anglophone philosopher of the late 20th century, David Lewis. Lewis was interested in the following puzzle.

On the one hand, it seems that languages are in some way conventional systems. It isn't a rule of nature that 'dog' will be the word for canines. After all, if it was a law, then it would hold in Paris as well as in London, and it doesn't. So it looks like it must be some kind of social arrangement that produces language - what else could it be?

On the other hand, it looks like it couldn't really be a social arrangement. After all, arrangements have to be made, and they are typically made in language. So if language is a convention, it must come after this arrangement was made. And that's impossible, since the arrangement requires language.

Lewis argued that this second argument, about the impossibility of formulating the agreement prior to language, was no good. He argued that conventions don't need to be anything like agreements. Rather, they can be equilibrium solutions to coordination games. All it takes for there to be a convention is that people play their part in an equilibrium, and they do so because it is in their self-interest to do this given that the equilibrium exists. The causal history of the equilibrium is irrelevant - it might have been a pure accident when it arose, but once it comes into place, it is a convention if people follow it because they have reason to follow it as long as everyone else does.

As well as this somewhat reductive account of what a convention is, Lewis popularised the 2-by-2-by-2 signalling game, as a model for how we might think about situations where these conventions come about. I'm not sure if he was the first to do this. (The work of Lewis's I'm talking about is in his PhD thesis, that became his 1969 book *Convention*, so what I'm not sure about is how much these games were talked about prior to 1969.) Lewis explicitly draws on the work by the economist Thomas Schelling, and Schelling talked about other coordination games. But in the current literature (or at least the philosophy part of it!) the credit normally goes to Lewis.

Reasons to be Sceptical of a Game-Theoretic Treatment

All that said, I'm a little sceptical that these general pictures about signalling can tell us much about the development of human language. The picture is that language is a solution to a coordination game, and that people play it because, I guess, they are good at detecting and continuing with solutions to coordination games. If that's right, then we have to assume that humans are good at either:

1. Computing the optimal solutions to games like these; or
2. Detecting and following regular social practices that are socially beneficial.

And while plenty of humans are actually good at these things, a lot of humans are not.

But, and this is the really surprising thing, humans are unbelievably good at picking up the language in their immediate vicinity. It's not true that 100% of humans develop competence in the local language, but it's stunningly close to 100%. Totally deaf people are an exception, and people with very severe learning disabilities are sometimes exceptions as well, but it's striking how few exceptions there are. Let's say, conservatively, that 99% of humans end up picking up the local language - at least to a level where one can get by. (I'm talking about spoken language here - for most of human history a small percentage of the population could read and write - but almost all can talk.)

It's worth thinking about what a bizarre fact this is. Try to think of any other practice that's as intellectually demanding as carrying out a conversation in the local language, and ask what percentage of the population are able to carry it out. Or think of any other beneficial social practice, and ask what percentage of the population go along with it. The answers will be well under 99%.

Of course, a lot of people are far from optimal users of language. But what I mean that the vast majority of humans can do is (a) parse utterances using common words in the local dialect, and (b) produce sentences that don't involve gratuitous grammatical violations. By grammar here, I don't mean Strunk & White rules. I mean that you just don't see vast numbers of humans producing sentences like "I are happy", or "You am tired". And it's really staggering how good people get at parsing the local language. Think how much study of Japanese it would take to be as good at processing everyday utterances in Japanese as virtually any three-year-old in Japan. Or how much French you'd have to study to be as good at correctly gendering the household furniture as pretty much any four-year-old in Paris. These are hard problems, and over 99% of kids figure them out somehow.

So I don't think we understand language in virtue of applying our general intelligence, or our general sociablility, to a coordination problem. We know what people are like when they apply general intelligence, or general sociability, and failures are really frequent. But failures at picking up the language, and conforming to its general rules, are really rare. This has suggested to many people that language must be associated with a special part of the brain, one that is designed to let us understand and produce sentences of a local language. (This idea, that language is associated with an innate, special purpose, system is often associated with the work of Noam Chomsky, though it's now a very widely held view.)

Now maybe we could still apply these game theoretic considerations 'one level up'. Maybe the reason we evolved a special purpose language system is because having such a system is an evolutionarily stable strategy. On this picture, it's not that we as individuals are trying (and succeeding) to solve a coordination problem. Rather, it's that we are 'programmed' to get to the solution instinctively, and the reason we are programmed this way rather than some other way is that this programming is part of a stable equilibrium. Maybe - but we should remember that language is very special, and that it is unlikely that general purpose reasoning will tell us just how it works.