Signals

Philosophy 444

2/14/23

Signalling Games

Here is the basic idea of a signalling game.

- There are two players, a sender and a receiver.
- Nature reveals some information to sender. (Or, if you want to make the game symmetric, nature chooses one of the two players to reveal information to.)
- Sometimes (especially in econonomic applications) we'll call this sender's **type**.
- Sender sends a signal that receiver can see.
- Receiver chooses an action that has a payoff to each player.

We'll start by considering **cooperative** signalling games. These are games where the players get the same payoff in every situation. Intuitively, they are ones where the players want to share information because they are in a joint venture.

To use a famous example, consider Robert Newman, the sexton of the Old North Church in Boston during the Revolutionary War. Part of his job (as a revolutionary) was to keep watch for what the Regulars were doing, and put signals in the window of the church to signal what they were doing. As the poem says, the signal was "One if by land, two if by sea". That is, he'd put out one lamp if they were coming by land, two if they were coming by sea. (Actually water; actually the Charles river. But that's not as poetic.) The other revolutionaries would then take suitable action.

So here Newman is the sender. Nature (well actually the British army) has revealed some information to him. (Inadvertently in this case.) Other people don't know this information. But they do know the signal he sends; they can see the number of lamps in the window. And they have a plan for what to do with each thing they see.

Now in reality, the plan they have is a good one. That's because they have coordinated in advance on what to do. But what if they can't coordinate. The game in this loose form has any number of Nash equilibria.

- It has **separating equilibria** where the townsfolk take different actions on seeing the different signals, and things work well for hearers and sender.
- It has **pooling equilibria** where Newman puts up the same signal come what may, and the townsfolk do the same thing no matter what he does.
- And it has **babbling equilibia** where Newman chooses randomly what to do, and the townsfolk ignore him.

But only the separating equilibria are ESS. There are two of these - we could have had one if by sea, two if by land. But apart from these two Nash equilibria, the other Nash equilibria are not evolutionarily stable.

This matters for two big debates

1. What is the relationship between human language and convention?
2. How do signalling systems evolve in non-human animals without agreement, or in most cases the capacity to do anything like make an agreement, as to what the signals will mean?

These are variants of the same question. How do you ever get a signalling system started? The answer being implicitly suggested here is that you randomly try a bunch of different strategies for navigating the world, keep the ones that work and not the ones that don't work (either intentionally or through natural selection), and eventually you'll end up signalling.

Signalling as Language

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

[1] A lot of this section of the 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.

> 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".

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

	S1	S2	S3	S4
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. [2]

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

[2] The main place to look for more information about this is Lewis's 1969 book *Convention*, which grew out of his doctoral thesis. It introduced many notions that became important in game theory, though in several cases Lewis was not the causal origin of the ideas; they had to be independently rediscovered by economists before they were widely adopted.

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. 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.

Non Cooperative Signalling

- Two states of the world - sender knows them, receiver doesn't. Call them 'High' and 'Low'. (Or H/L.) Assume for now that High is marginally more probable than Low.
- Two possible signals/messages that sender can send - call them 'Difficult' and 'Easy' for reasons that we will get to. (Or D/E).
- Two possible actions receiver can take - call them Risky and Safe (Or R/S).

In words, here are the payoffs:

- Sender pays 0 to perform Easy, but a cost $c > 0$ to perform Difficult. This is subtracted from whatever their ultimate payout is. (We will complicate this clause in a bit.)
- Receiver gets a payout of 0 for doing Safe.
- If Receiver plays Risky, they get 1 if High, and -1 if Low.
- Sender gets 1 (minus whatever cost they incurred at round 1) if Risky, and 0 (minus whatever cost they incurred at round 1) if Low.

In table form, here are the payouts. First, here are the payouts for High.

	Risky	Safe
Difficult	$1 - c, 1$	$-c, 0$
Easy	$1, 1$	$0, 0$

Now, here are the payouts for Low.

	Risky	Safe
Difficult	$1 - c, -1$	$-c, 0$
Easy	$1, -1$	$0, 0$

Let's try to work through what the equilibria are for different values of c.

- First, assume $c < 1$.
- Assume that Receiver will do different things if Difficult or Easy.
- Then Sender will do whatever makes Receiver do Risky, whether it is D or E. So no separating equilibrium.
- Now assume Receiver will do the same thing if Difficult or Easy.
- Then of course Sender will do Easy and get that payout, whatever it is.
- Again, no separating equilibrium.

What does produce a separating equilibrium is if the cost is different in High and Low. Change the 'penalty' for playing Difficult so that in High, the penalty is $c_1 < 1$, and in Low, the penalty is $c_2 > 1$. Now we get the following equilibrium.

- Sender does Difficult if High, Easy if Low.
- Receiver does Risky if Difficult, Safe if Easy.

(Aside: You also get one other really weird equilibrium where

- Sender always does Easy.
- Receiver does Risky if Easy, Safe if Difficult.

This is mathematically interesting - and hence interesting to me - because it is not just a Nash equilibrium and subgame perfect, but also satisfies all the extra criteria developed in chapters 11 and 12 to rule out intuitively absurd equilibria like this one. But it is hard to see how it has any real-world relevance - it doesn't look like it could naturally evolve, for example - and I'll ignore it from here on.)

So we get a separating equilibrium. And maybe we get a model of some fascinating real-world things. In most of these cases, Sender has something like a continuum of choices from Easy to Difficult, and Receiver has a continuum from Risky to Safe. But it arguably helps to look at the binary case first, and maybe we can generalise that to the real-world example.

Two caveats before we start the examples.

In practice, biologists seem happy to use this procedure - think about the binary model and then generalise to the continuous case - while economists prefer to start with the continuous case. My intuitions are normally with the economists, but here I'm acting like a biologist.

And these are possible models of what we see. In every case, I'm going to eventually raise worries for the model. But I want to have them on the table.

Example One: Tail-Feathers

Male peacocks have very colorful tails. On the face of it, this doesn't look like it serves any purpose in either collecting food or avoiding becoming food. (Quite the opposite in fact.) But maybe we should think of it as a move in a signalling game.

- Sender is the male, choosing whether to have a normal tail (Easy) or a colourful tail (Difficult). 'Choosing' here is misleading - it's less misleading to say their genes choose.
- Sender is either Strong (that's High) or Weak (that's Low).
- It's resource-intensive to produce (and preserve) a colorful tail, but it's more costly for Weak than for Strong peacocks.
- Receiver is a female, choosing a mate. They prefer Strong to Weak - since they want better genes for their children. (Again, it's hard to say this is what the individual female wants - better to say there are evolutionary advantages to acting as if that's what she wants.)
- So perhaps an equilibrium is Strong have colorful tails, Weak don't, and females who have a choice prefer males with colorful tails.

Example Two: Stotting

Stotting is where a quadruped leaps into the air, with legs relatively stiff. Stotting is common among young animals in various species. But the really odd thing is that among some gazelles, it only happens when a predator is nearby. And why they do this is a bit of a mystery. It doesn't seem that efficient as a means of propulsion. And revealing one's location this dramatically doesn't seem like a good tactic in predator avoidance. But maybe it's a move in a signalling game. In this game, the payouts are slightly changed. The gazelle is the sender, and the predator is the receiver, and the predator's payoffs are going to be different from the standard game.

- High in this case is that the gazelle is strong (High means things are good from the predator's perspective.) Low is that the gazelle is weak.
- Stotting is Difficult; Not-stotting is Easy.
- Chasing this particular gazelle is Risky; leaving it is Safe.
- But here we change the payoffs. Here receiver/predator gets -1 for doing Risky in High, and 1 for doing Risky in Safe. Otherwise the game is the same.

Again, there is a separating equilibria.

- Gazelle stotts if and only if they are strong.
- Predator chases if and only if they don't see stotting.

And so there is an advantage to stotting - you don't get chased - even though holding fixed the state of the world and the behavior of the predator, stotting is a cost with no benefit. It doesn't help you get away - it helps the predator not choose to attack.

Example Three: University

According to recent research from the San Francisco Federal Reserve, here are the average hourly wages for Americans by educational level as of 2015. (I don't think the numbers have changed much since.) I've also added a column for what percentage of the workforce each of these groups are. Source: https://www.frbsf.org/economic-research/files/wp2016-17.pdf

Education	Wage	Ratio
No degree	$13.56	7.7%
High school degree	$17.98	25.6%
Some college	$21.59	27.8%
Undergrad degree	$30.93	24.7%
Graduate degree	$39.48	14.3%

What could explain the fact that college graduates earn almost 75% more per hour than high school graduates? There are two obvious possibilities.

1. Universities impart lots of valuable skills, and employers are rationally responding to this value we add by paying more for more valuable employees.
2. It's a selection effect - the people who come to college were more valuable before they came here, and employees are rationally responding to that underlying fact.

These aren't exclusive, but let's pretend for now we're going to assume one of them is decisive, and we're trying to figure out which it is. There are two things missing in explanation 2.

First, why do all these smart people choose to go to college? On the one hand, it is mostly fun. On the other hand, it's expensive, and there are a lot of other fun things you could do with the money. If it's all about the climbing walls, you could join the fanciest gym in the country for a fraction of the cost. If it's about meeting new people, you could go backpacking around Europe. If it's about intellectual stimulation, you could take a gap year or four and spend your days reading and listening to educational podcasts.

Second, why do employers look for college degrees as the signal of who they will pay high wages to? Why don't they simply ask to see your offer letters? If it's just a selection effect, then you can see who has been selected in as soon as the offer letters go out - and that should be enough to make employers happy. But it's not - they want degrees not just offer letters. (This seems really obvious, but I think it's kind of a striking fact about the modern world, and one that doesn't get enough attention in a lot of debates.)

So we need a model that explains why we don't see, for example, valuable employees taking their offer letters and backpacking around Europe with Kindles and podcasts for their spare time. Spence's signalling model provides such an explanation.

- Sender is the college graduate. They are either High - i.e., valuable to employers, or Low - not so valuable.
- Receiver is the high wage employer. They can do the Risky thing - hire this person - or the Safe thing - not hire them.
- Going to college is Difficult. But - and this is the crucial thing - on this model it is more Difficult for Low than for High. All those calculus classes are really unpleasant. But they are more unpleasant for Low - so much so that $c_2 > 1$.
- So the separating equilibrium is High goes to college and Low hits the beach/workforce. Then high-wage employers hire college graduates only.
- And all this happens even though college is a pure cost to everyone who goes there. Employer doesn't get any reward for hiring graduates - they get same reward for hiring High grads as High non-grads. And holding all else fixed, every young person is better off skipping college than going.

Four questions for you to discuss.

1. What aspects of this model seem to resemble the world you (as in literally you personally) find yourself in?
2. What aspects of it seem to differ from the world in significant ways?
3. What empirical data would make you think this model was right in some important way?
4. What empirical data would make you think this model was wrong in some important way?

Honest Signalling

Big Themes

- There are other signalling models than the Spence model. You can think that tail feathers or stotting or college attendance is signalling without thinking this is the right model.
- The Spence model says that non-signallers **can** signal but **won't**. We should also think about models where non-signallers **would** signal but **can't**.

Three Big Questions

1. Why does receiver take the signal seriously?
2. Why does signaller send signal? In particular, why do they send this signal? Why don't signallers collectively organise a cheaper signal?
3. Why don't non-signallers send signal?

Four Questions from Last Time

1. What aspects of this model seem to resemble the world you (as in literally you personally) find yourself in?
2. What aspects of it seem to differ from the world in significant ways?
3. What empirical data would make you think this model was right in some important way?
4. What empirical data would make you think this model was wrong in some important way?

My Answers

1. The basic structure seems plausible. It isn't obvious how what we do here makes you a more valuable employee. It might make you better citizens, but employers don't care about that. And calculus class really is less pleasant for people who won't be as valuable.
2. The wage premium is so high that it's hard to believe $c_2 > 1$. The same is true for the tail feathers and stotting examples. The payouts to Difficult are really really high, and calculus class isn't that differentially unpleasant.
3. Restricting things to data I know exist - one thing that supports the model is the 'sheepskin effect'. If you divide the 'some college' group by how much college they got, the skills hypotheis would predict that the more college you did, the higher your wage premium. That might be approximately true. But it would also predict that if you're one course from graduation, you would get 95% or more of the wage premium. And that's wildly false.
4. One thing that's trouble for the model is that the wage premium is really high among older workers. To make that work, you need one of a few implausible things. One possibility is that you need employers who are so unobservant that they can't tell the valuable workers from the not valuable workers after years and years of work, so they still have to rely on degrees as a signalling device.

Another Hypothesis

So far we've assumed everyone can do Difficult or Easy, but it is more costly for Low than High. Maybe we should drop that assumption. Here is another kind of signalling model that we could consider.

- The basic structure of sender, receiver, signal and receiver actions are the same.
- Now the cost of Difficult is the same for High and Low.
- But now Sender isn't in full control of their actions.
- If they choose Easy, then Easy happens.
- But if they choose Difficult, then they have to pay the cost c, but what happens, and what Receiver sees, is Difficult with probability p, and Easy with probability $1 - p$.
- And the probability is high for High and low for Low.
- Again, we get a separating equilibrium.

- Receiver does Risky if they see High, and Safe if they see Low. (Or the other way around in the stotting game.)
- And the expected payouts are such that High should take the chance and do Difficult, while Low should not.
- At the extreme, the probability of success is 1 for High and 0 for Low, but the model works even with much more balanced probabilities.

This is sometimes called the **Honest Signalling** model, or the **Indexical Signalling** model, as opposed to the **Handicap Principle** model we started with.

Mixing the Hypotheses

Maybe for some Senders, they have a choice about what costs to incur, with the more costs they incur increasing their probability of success.

- This is easier to see in the college case. Imagine a person who has the skills to finish a college degree, but doesn't have the skills to finish while holding down a 40 hour job, and it would be really costly to give up the 40 hour job.
- For that person, going to college and keeping the job might not be worth it - it would be like doing the low-probability Difficult signal, which usually just results in paying a cost and getting no return.
- But going to college and giving up the job might not be worth it either. If the cost c isn't just the tuition cost, but the money foresaken from the job, and perhaps the interest on the loans taken out to cover that money, then perhaps the cost is higher than the premium.

In general, we might want to be a little sceptical that there is clean line between cases where a player chooses not to send a costly signal, and cases where that player doesn't have the ability to (reliably) send that signal. Maybe the signal success probability is a function of the costs incurred, and there is no level of costs they can justify spending.

For More information

For stats on the college wage premium over time, see

- https://fredblog.stlouisfed.org/2018/07/is-college-still-worth-it/

For information on the college wealth premium, plus stats on the demographics of both the wage premium and the wealth premium, see

- https://www.stlouisfed.org/~/media/files/pdfs/hfs/is-college-worth-it/emmons_symposium.pdf?la=en