PREPRINT 11 June 2018

PROPOSITIONAL CONTENT in SIGNALS

(for Special Issue *Studies in the History and Philosophy of Science C*)

Brian Skyrms and Jeffrey A. Barrett[1]

"Truth lives on in the midst of deception." -Schiller

Introduction

We all think that humans are animals, that human language is a sophisticated form of animal signaling, and that it arises spontaneously due to natural processes. From a naturalistic perspective, what is fundamental is what is common to signaling throughout the biological world -- the transfer of information. As Fred Dretske put it in 1981, "In the beginning was information, the word came later." There is a practice of signaling with information transfer that settles into some sort of a pattern, and eventually what we call meaning or propositional content crystallizes out.

The place to start is to study the evolution, biological or cultural, of information transfer in its most simple and tractable forms. But once this is done, naturalists need first to move to evolution of more complex forms of information transfer. And then to an account of the crystallization process that gives us meaning.

There often two kinds of information in the same signal: (1) information about the state of the world that the signaler observes (2) information about the act that the receiver will perform on receiving the signal. [See Millikan (1984), Harms (2004), Skyrms (2010)]. The meaning that crystalizes out about the states, we will single out here as propositional meaning.

The simplest account that comes to mind will not quite do, as a general account of this kind of meaning, although it may be close to the mark in especially favorable situations. That account is that evolution leads to, or approximates. an equilibrium in usage where prefect information about some state (or set of states) of the world is transferred, and that this information gives us the propositional content of a signal. If usage is at, or approximately at, a separating equilibrium in a Lewis signaling game, this simple account seems fine.

But there are lots of cases of established usage in which information transfer is at odds with what we take as literal meaning. The speaker may say something that is literally false. (The

[1] We would like to thank Manolo Martinez, Jonathan Birch, Peter Godfrey-Smith, and Nickolas Shea for helpful comments.

individual or individuals hearing it may or may not be alive to the possibility that it is literally false.) The propositional meaning (according to which it is false) is at variance with the information transferred. For instance, one might think of the opening gambits of the used car salesman, the realtor, or the seller in the bazaar. The information contained in the utterance is roughly "the salesman's opening claim is such-and-such," but the content is taken as propositional, and it may be literally false. Think of the little boy who cried wolf.

Humans studying animal signaling transfer this distinction to animals. Consider birds that make false alarm calls when they find food to scatter the flock so that they can eat more. Or consider false mating signals sent by females of one species so that they can lure males of another and have them for dinner. These are not rare occurrences, but rather relatively frequent.[2] We cannot just dismiss them as the rare sort of out-of-equilibrium behavior covered by the qualifier "approximate". Biologists describe these as cases of deception. The content is taken as "there is a predator" or "I am a sexually receptive female of your species", while the information just has to do with what is correlated with the signal being sent. If we took the information to be the meaning, "predator present or I'm fooling you to get more food", "receptive same species female or hungry predator" then there would be no falsehood possible. Some might take this point of view, but if one took it with respect to human conversation it would lead to the same conclusion. This would simply obliterate useful distinctions.

Biologists and philosophers, some represented here, have developed naturalistic accounts that preserve the distinction between content and information. We have been moved to think about this issue by the recent work of Peter Godfrey-Smith (2012) and Jonathan Birch (2014). There is a lot of commonality in these proposals, but there are significant differences as well. We will join the crowd as kindred spirits with a somewhat different proposal. We will operate, as others have, within a signaling game framework. These are all, in game theory terminology, games of incomplete information. But in cases of prime interest here, as we shall explain, the information can be thought of as incomplete is several dimensions.

Signaling Games

The well-known signaling games of David Lewis (1969) provide models of information transfer from sender to receiver via signals in a benign situation in which the players have *common interests*. Suppose a husband phones wife and asks whether it is raining where he plans to meet her. If she says rain, he brings and umbrella; if not he doesn't. It is to both their advantage of he gets it right. More abstractly, a sender observes a state of the environment, and sends a signal; a receiver observes a signal and chooses an act. If the receiver chooses the act that "matches" the state both are paid off; otherwise neither.

---

[2] In some species, the majority of alarm calls given are false. See Searcy and Nowicki 65-68.

Payoffs are not solely determined by the combination of acts of sender and receiver. The state observed -- in game theory terminology the "type" of the sender -- is also relevant. One type of sender sees rain; another sees sunshine. We can conceptualize the interaction thus: The environment picks a state with certain probability, the sender observes the state (her own "type"), and sends a signal according to her signaling strategy. Then the receiver observes the signal and chooses an act contingent on the signal, according to his strategy for reacting.

In such repeated games of common interest, it is not implausible to assume that such interactions settle into a game-theoretic equilibrium. Then the content of the signal may just be read off the equilibrium. "Rain" means rain; "sunshine" means sunshine.

The problematic cases of the previous section are not games of common interest. The vendor in the bazaar who says "This is a genuine Rolex." or "This is a genuine Louis Vuitton handbag." does not share common interest with the person being addressed. The "femme fatale" firefly of genus *Photuris* who sends a mating signal to a male *Photinus* does not share common interest with the male that she will, if he is attracted, eat for dinner.[3]

In models of these interactions as games, the content cannot simply be read off the equilibrium. There is typically an equilibrium that includes both instances of what we would like to think of as honest signaling and instances of what we would like to think of as dishonest signaling. The sender in the bazaar is not the only sender of the "Rolex" signal, the used car salesman is not the only one using "mechanically sound', the *Photinus* mating signal is also sent by *Photinus* females ready to mate. "Honest" encounters and "dishonest" encounters occur in proportions adequate to maintain an equilibrium.

Thus, the operative type in the general game consists of two things: (1) what the sender observes and (2) what we might call the *interaction context*. The environment (or "nature") picks the type, which is a pair <state observed, interaction context> with a certain frequency and signaling interactions evolve. Such evolution may settle into an equilibrium (or quasi-equilibrium) of information transfer. Typically, such equilibria are mixtures of intuitively "honest" and "dishonest" signaling. To say this, we need content that is not simply read off the equilibrium.

We can think of the environment picking the type as a 2-stage process. First it determines the context, then it determines the state that the sender observes. Some contexts may be contexts of common interest. That is to say, after the selection of some contexts, the subgame that we are left with is essentially a Lewis signaling game, as shown in figure 1. Our suggestion is that in these cases content is to be read off an equilibrium in this signaling game that we get by restricting to contexts of common interest.

---

[3] Lewis and Cratsley (2008).

Notice that we have not drawn in the information sets in figure 1. That is because there are special cases. In all cases, we assume that the sender observes the state, and the receiver does not, just as in Lewis' model. In all cases, we assume that the sender observes the context. But we have the case where the receiver does not observe the context and a second case where the receiver does. In case the receiver also observes the context, we have a pure Lewis signaling game in the common interest context, and we should expect no information transmission in the opposed interest context. This is like our used car example. Where the receiver does not observe the context, we have the possibility of deception, as with the fireflies. Many real phenomena may be intermediate cases where the receiver may observe the context imperfectly, and dealing with these may be challenging on both a theoretical and empirical level.[4]

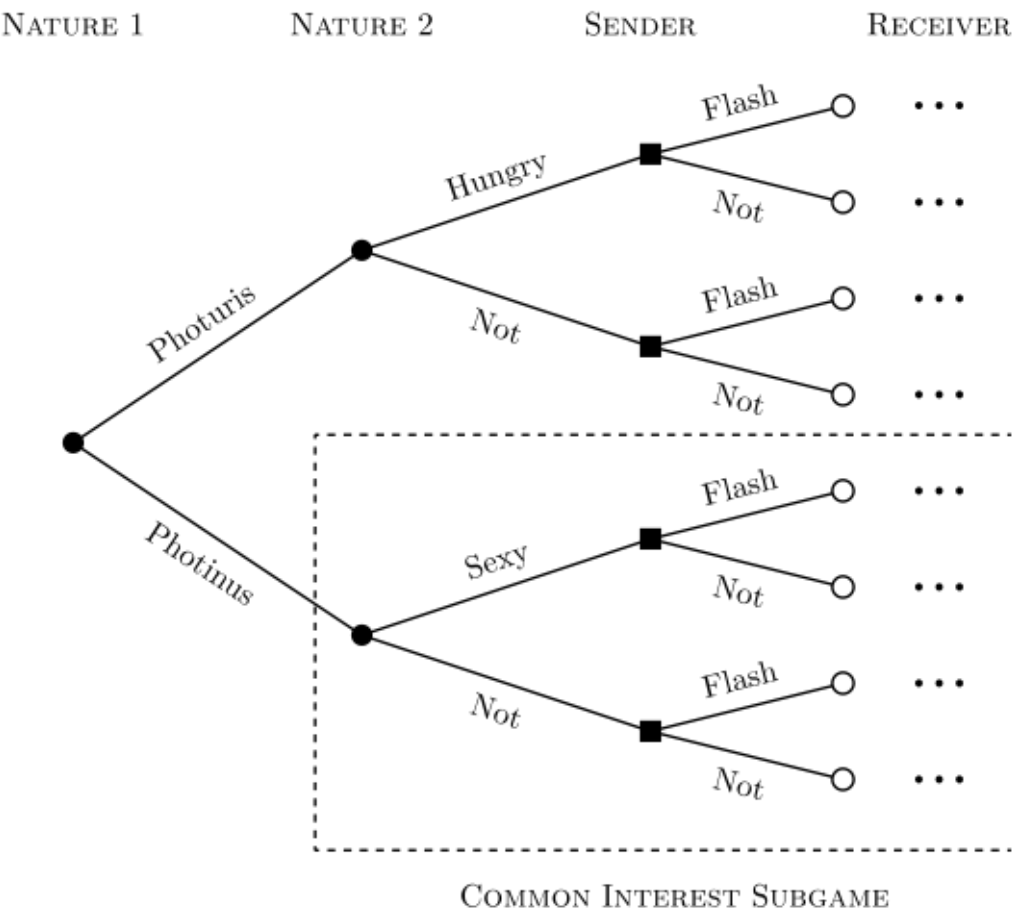

figure 1: Common Interest Subgame in Firefly Signaling

[4] See Wheeler and Hammerschmidt (2013)

"Wolf" does not get its meaning from contexts that include the little boy crying wolf. This would be true even if crying wolf became more common than in the story. "Louis Vuitton" does not get its meaning from the vendors selling fake plastic handbags, although they may account for most of the usage. What we take as the content of the signal is lifted from usage in contexts of common interest. It is from these patterns of usage in contexts of common interest that meaning crystallizes and becomes separable from the pragmatics of information transfer. We submit that this is a rule tacitly used by those who codify patterns of usage in dictionaries. (All the more so because individual words appear in a variety of different sentences.)

It is therefore quite natural for biologists who study patterns of information transfer in non-human animals to use the same rule. Photuris is considered to be sending a signal meaning "I am a receptive Photinus female" rather than "I am just the sort of thing that sends this signal," and thus the signal is classified as deceptive. The forked-tail Drongo making an alarm call when no predator is present to steal food[5] is considered to be sending a signal that means "danger from predator" rather than "either danger from predator or I want to eat your lunch," and thus the signal is classified as deceptive.

We do not, however, wish to confine deceptive signaling to cases where there is a natural propositional content available. Rather, we prefer a broader approach to deception along the lines suggested in Skyrms (2010). On this view, a signal is deceptive if it carries misinformation and is consistently in the interest of the sender and to the detriment of the receiver given the payoffs of the game. There is now a literature on this functional approach to deception including recent papers by Martinez (2015) and Fallis and Lewis (2017).

Interaction with Jonathan Birch's Proposal

Jonathan Birch (2014) put forward a different proposal for propositional content that seems quite different, but there are affinities between his idea and ours. His proposal is that the meaning is the information transmitted by the signal in the separating equilibrium closest to actual behavior. "Closest" is glossed in different ways depending on whether there is a separating equilibrium in the game at all, or one needs to move to a modified game to find one. In the first case, distance is just the Euclidian distance in the space of probabilities of strategies. In the second it is a distance in a space of parameters in the game.

We first ask how this works in a simple Lewis signaling game. Here Birch's proposal makes precise what to do in the case we glided over previously by saying that the population behavior was at an "approximate equilibrium", and that the meaning was "read off" the equilibrium. This seems correct. The meaning is not the information transferred in the

---

[5] Flower (2011).

approximation, but rather that transferred in the pure equilibrium that it approximates. This makes room for infrequent mistakes or other deviations.

There are, however differences. In Lewis signaling games, suppose that we are not close to a separating equilibrium.[6] Suppose instead, that we approximate what is called a partial pooling equilibrium. For instance, suppose that there are three states, three signals and three acts, with the states equiprobable. The sender always sends signal 1 in states 1 and 2, and sometimes sends signal 2, sometimes signal 3 in state 3. The receiver sometimes does the act 1 appropriate for state 1 when seeing signal 1, sometimes the act appropriate for state 2. When seeing signal 2 and 3 the receiver always does the acts appropriate to state 3, as shown in figure 2.

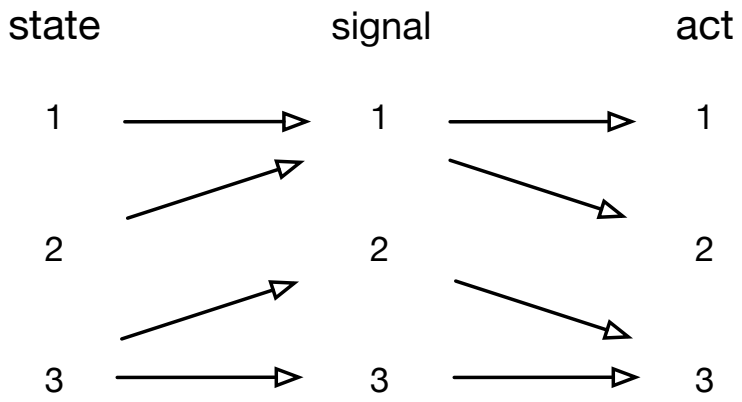

figure 2: Partial Pooling

Figure 2 represents not just a single partial pooling equilibrium, but a whole class of them, depending on the proportions with which senders and receivers mix between signals 2 and 3, and acts 1 and 2 respectively. Suppose the sender, on seeing state 3, 1% of the time sends signal 2 and 99% of the time sends signal 3. And suppose that the receiver on seeing signal 1, 99% of the time does act 1 and 1% of the time does act 2. Then, evidently, the closest separating equilibrium is the one where both sender and receiver switch their behavior with respect to the rarely used signal 2. That is the equilibrium in which the sender always sends the signal whose index (in the numbering of figure 2) matches the state and receiver always chooses the act whose index matches the signal. This remains true if instead of 99%-1% in the partial pooling equilibrium, we have 51%-49%; closest does not mean close. Would we then want to say that

---

[6] Since Lewis signaling games are common interest there will always be a separating equilibrium for *n* x *n* x *n* games.

signal 2 *means* state 2 in this case, even though it is never sent in state 2, but often sent in state 3?

Since we are already in a game of common interest and since the game is in a state of equilibrium, in contrast to Birch, our proposal would take the meaning from the equilibrium. So our proposal would say that signal 2 means state 3.

In games of mixed interests Birch looks for the closest game in parameter space with a separating equilibrium, and the closest separating equilibrium in that game. Our proposal looks for a context such that individuals in that context have common interest, and looks at behavior in that subgame. There is no guaranty that such a context exists. When it does, our proposal gives one sense of an appropriately close game. The closest game with a separating equilibrium may often be the subgame that we look at, though this is not guaranteed. But even if they agree on an appropriately close game of common interest the two proposals may disagree as above.

Equilibrium behavior in the original mixed interest game or may not may not be an equilibrium in the common interest subgame. Predators may be bad enough that birds scatter on hearing an alarm call; sex drive may be strong enough that male fireflies throw caution to the winds. In these case cases, restriction to the common interest game gives us a separating equilibrium. Here Birch's proposal and ours may well come to the same thing.

But in other cases, behavior in the common interest subgame may fall short of separating equilibrium behavior. The cautious used car buyer may be somewhat suspicious of even his long-time, thoroughly honest mechanic. The villagers may become a little blasé about shouts of “wolf!" In these cases, Birch's idea is a useful supplement to ours. "Wolf" still means wolf. In the case of behavior close to an equilibrium in a game restricted to contexts of common interest, we can use information transmitted in that equilibrium to give us propositional content.

The main differences between the two proposals are (1) our restriction to contexts of patterns of observed behavior in sub-contexts of common interest and (2) our use of "close" rather than "closest". The conditions for applying our proposal may not obtain, in which case it has nothing to say about meaning.

Interaction with Godfrey-Smith's Proposal

Peter Godfrey-Smith (2012) proposes a revised and updated form of Ruth Millikan's teleosemantics. Millikan suggested that the meaning in a signal is to be found in the reason why it evolved. Godfrey-Smith suggests instead that we look at the reason or reasons why evolution *maintains* a signal in use. The shift seems necessary if we are to properly account for the fact that signals can change their meaning over time. In the process of cultural evolution this happens over a few generations. In genetic evolution, it takes longer, but it still happens. Historical

reasons may be more relevant to your grandfather's, or your grandfather species', meaning than to yours.

More recently, Shea, Godfrey-Smith and Cao (2017) have developed a richer and more detailed version of these ideas within the framework of signaling games. "Maintaining reasons" are made precise in a *functional content vector*, that takes a place beside the *information vector* carried by a signal. In appropriate cases, a *narrative summary* is available, and this is what is closest to what we have been calling the "propositional content" of the signal. The *narrative summary* of a vector is the disjunction of the non-zero coordinates of the vector. Both information vectors and content vectors have non-trivial narrative summaries when they have some zero entries.

The content vector is a kind of summary of the benefit received by the players from signaling in each state. Zeros correspond to no benefit, or even negative benefit compared to a no-signaling baseline. It is useful to have such a summary of benefit from signaling. The basic idea is that benefit is the reason that the signaling is maintained, and that this general idea is made precise by way of the content vector. There are technicalities involved in the definition of the content vector in Shea, Godfrey-Smith and Cao that reflect decisions that could have been different. Variations on their basic idea are possible, and some of these might also be of interest.

If we compare the *narrative summaries* of Shea, Godfrey-Smith and Cao with propositional contents according to Birch and to us, one salient difference is that we use the equilibrium concept as a way of taking both the interests of senders and receivers into account whereas they take account of these interests in a different way. It is then possible for signals to have non-trivial narrative summaries when the system is out-of-equilibrium and even far from equilibrium. For instance, signal 1 may be sent in state 1 and only in state 1, the receiver may, on seeing this signal, do the act that is best for state 1, and this may be very good for both sender and receiver (and it may be very bad for both for the receiver to do otherwise in state 1) even though the system is far from equilibrium in other states, and may indeed never approach equilibrium because of cycles or chaotic dynamics.

But the relevant equilibria for Birch and for us are not always actual ones in the actual game, but rather related ones. Everyone agrees when we are at a separating equilibrium in a common interest Lewis signaling game. If we are at the partial pooling equilibrium in the Lewis signaling game shown in figure 2, we will agree with Shea, Godfrey-Smith and Cao that signal 1 "means" state 1 or state 2, and that both signals 2 and 3 "mean" state 3. Birch, if we am correct above, will look for the closest separating equilibrium and may come to a different conclusion.

What about Photuris and Photinus? Our account says that when Photuris sends the Photinus mating code, the propositional content of her signal is best expressed as "I am a sexually receptive Photinus, ready to mate".

We have a game with 4 states:

S1: Photinus, ready to mate
S2: Photinus, not interested
S3: Photuris, ready to eat
S4: Photuris, not interested

(States 2 and 4 are much more common than states 1 and 3)

There are 2 signals in play:

M1: The Photinus mating flash pattern.
M2: No flash (the null signal)

The acts are:

A1: Approach
S2: Don't

Nature's strategy profile:
M1 in S1 and S3, M2 otherwise, for the sender
A1 if M1, A2 otherwise, for the receiver.

What is the baseline? We suppose seeing no signal Photinus would not approach, because with no signal which blade of grass to approach? Supposing that is correct, content vectors for signal M1 for sender and receiver disagree.

Sender <1, 0, 1, 0>
Receiver < 1, 0, 0, 0>

We then take the minimum at the points of disagreement, and get <1, 0 0 0>, as an overall content vector.

If so the content vector analysis leads to the same gloss on the meaning of M1. It is "I am Photinus, ready to mate."

Commonalities

We all agree that content arises from information transfer. Content is information that has become ritualized[7] and decoupled[8] from the relevant contexts in which content and information were the same. Once this happens to a signal, its content may diverge from the information that it carries, as in the examples discussed here. Semantics is born from pragmatics, but then they become separate.

There are some detailed differences in these accounts. We do not think that there necessarily has to be one "right" account, down to the last detail. There is also a large class of cases in which the various proposals regarding content do agree. This is, perhaps, enough to suggest that we are on the right track.[9]

Learning to Lie -- an example

The foregoing is all at a high level of generality, intended to cover a variety of situations and susceptible to dynamic analysis on multiple levels. Here we discuss one kind of learning dynamics for one specific case, in order to provide an example of how a propensity to send false signals in specific contexts may arise.

Wheeler (2009) describes how capuchin monkeys (Cebus apella nigritus) use terrestrial predator-associated alarm calls for the purpose of tactical deception. In the experiments he describes, some monkeys are observed to produce high-urgency alarm calls to cause their colleagues to flee so that they can steal food that they would not otherwise get.

The monkeys studied use three acoustically distinct predator alarm calls. The *bark* is used in response to aerial threats, and the *peep* and *hiccup* are used in response to terrestrial threats. The number and rate of hiccups a monkey produces are correlated to the degree of risk it faces.

---

[7] Barrett and Skyrms (2017).

[8] Decoupling here is used in a rather different sense from Sterelny (2003).

[9] We take the present account of the conception of propositional content to be an extension of that expressed in Skyrms (2010). In the chapter on information, Skyrms claims that the propositional content of a signal "can be read off of the informational content vector." It is essentially the disjunction of states that are not ruled out. This proposal was made in the context of standard Lewis signaling games, although a restriction to those games was not stated (and perhaps a wider application was implied). With the benefit of hindsight, and subsequent literature, we hereby makes that restriction. And with that in place, the position expressed here fits with Skyrms (2010) and can be thought of as an extension of that account.

Two or more hiccups in quick succession are taken to constitute a high-urgency terrestrial alarm call.

In order to study the monkeys' use of alarm calls for tactical deception, the experimenters placed banana pieces on feeding platforms as the capuchin group approached, then noted their use of high-urgency terrestrial alarm calls. Nearly every observed case of tactical deception involved a subordinate monkey trying to steal food from a more dominate colleague.[10] Deceptive alarm calls caused escape reactions in nearby dominate monkeys about 40% of the time. When effective, the deceptive alarm calls increased the feeding success of the deceiving monkey about 70% of the time. When the deception worked, the subordinate monkey would grab the banana pieces immediately after its colleague jumped from the feeding platform. The monkeys have evolved both meaningful alarm calls and the ability to use them to lie when an appropriate opportunity presents itself.

Here we consider how evolved signals might come to be used deceptively in the context of a hierarchical signaling game under simple reinforcement learning.[11] The model captures some of the salient aspects of the behavior of the capuchin monkeys. There are two stages, in each the agents play a different game.

In the first stage, the sender and receiver play a simple 2x2x2 Lewis signaling game. In this game, the sender observes the state of unbiased nature, then sends a signal. The receiver, who cannot observe the state of nature, then performs an action that is successful if and only if it matches the state. Here success means that the receiver *flees* if and only if the state is *predator*. If the receiver's act is successful, then the disposition that led to each agent's action on the current play of the game is reinforced.

In the second stage, the sender becomes sensitive to context and to the possibility of manipulating signals. We model this by introduce a module in the sender that can change the signal that the sender would otherwise send dependent on context. We call this module an *executive sender*.[12] This changes the game. The executive sender's action is sensitive to the natural context, either *business as usual* (if there is *no opportunity* to deceive or if there is a *predator* nearby) or *safe chance* to deceive (if there is an *opportunity* to deceive and *no predator* nearby). As his action, the executive sender may send the term that has evolved to represent the current state in the basic game (that is, he may perform the action *no lie*) or just send the term

---

[10] Given the empirical evidence from the experiments, one would expect to see a randomly-selected monkey using a high-urgency deceptive alarm call to get food on about 5% of the trials.

[11] Hierarchical games are discussed in Barrett, Skyrms, and Cochran (2018).

[12] See Barrett and Skyrms (2017) for an account of how a more complex game, like the second-stage hierarchical game we consider here, might evolve from simpler games by means of modular composition.

that has evolved to represent the state *predator* (which will be a *lie* if there is in fact no predator). See figure 3 for a picture of the second-stage game.

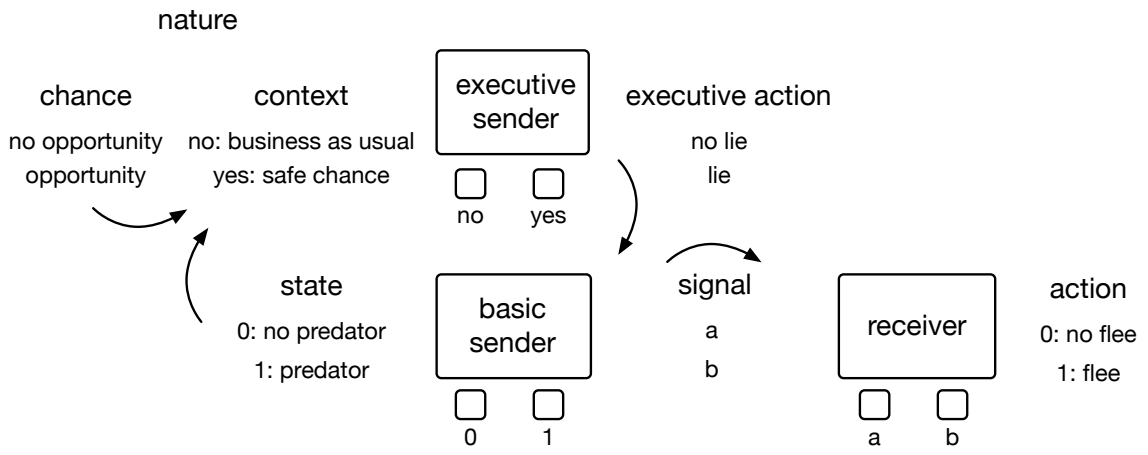

figure 3: a hierarchical model for the evolution of lying

We will assume that all three of the agents learn by *simple reinforcement*. In the simple first-stage game, one might imagine the sender with two urns, one labeled 0 for *no predator* and one labeled 1 for *predator*. At the beginning of the stage, each of these urns contains two balls *a* and *b*. When the sender sees the state, she draws a random ball from the corresponding urn and sends the signal indicated on that ball. The receiver also has two urns, one labeled *a* and one labeled *b*. Each of these urns contains two balls *no flee* and *flee*. When she sees the signal, she draws a random ball from the corresponding urn and performs that action. If the action matches the state, then the agents are successful and each returns the ball she drew to the urn from which she drew it and adds a ball of the same type to that urn; otherwise, the agents simply return the balls they drew to the urns from which they were drawn.[13]

The sender and receiver start by randomly signaling and acting. But this simple 2x2x2 signaling game with unbiased nature and simple reinforcement learning is guaranteed with probability one to evolve a signaling system where one term reliably indicates *predator* and produces the action *flee* and the other reliably indicates *no predator* and produces the action *no*

[13] The agents here do not face the risk of being eaten by a predator in this model. A model that considers agent survival would need to track both the evolution of agent types in a population from generation to generation and how each type learns within a generation.

flee.[14] Of course, there is no way to tell up front which term will ultimately mean what on a run of the basic signaling game.

The second-stage game begins when the sender and receiver have learned to signal reliably in the basic game. In this stage, the executive sender may learn to use the evolved signals to lie when the occasion allows. The executive sender has two urns, one labeled *business as usual* and one labeled *safe chance* to deceive. These urns represent the natural context. If there is an opportunity to deceive and there is no predator present on the current play, then the executive sender draws a ball from the *safe chance* urn; otherwise he draws from the *business as usual* urn. Each of these urns begins the second stage with one *no-lie* ball and one *lie* ball. If the executive sender draws a *no-lie* ball, he sends the receiver the signal that has to this point evolved to represent the current state of nature. This is the type of ball the basic sender would be likely to draw from the current context urn. But if the executive sender draws the *lie* ball, he simply sends the signal that has to this point evolved to indicate the presence of a *predator*. So if the current context is *safe chance*, then there is in fact no proximal predator and the signal is a lie.

The context *safe chance* in the second-stage game corresponds to a situation where a subordinate monkey has a chance to steal food from a colleague by falsely signaling the presence of a predator, and the context *business as usual* corresponds to a situation where everyone would be better off using the evolved signals in their usual senses. This suggests the following second-stage payoffs. If the context is *business as usual* and the receiver's action matches the current state, then both the executive and basic sender and the receiver are reinforced with one ball on what they did this play. This mirrors what would happen on a play of the first-stage game, but the action the executive sender took is also reinforced now. If the context is *safe chance* and the receiver *flees*, then both senders get reinforced with two balls on whatever they did this play. In this case the senders lied, they knew they were lying, and it worked. Finally, if the context is *safe chance* and the receiver does the action *no flee*, then the receiver's action is reinforced with one ball. Here the senders' attempted deception failed, and the receiver is rewarded for doing the right thing since there is in fact no predator.

Given an unbiased chance of a predator and an unbiased chance for deception, on simulation, the executive sender typically learns to lie using the signaling system that evolved in the first stage. Further, both the meaningful signaling system that evolved in the first-stage game and the executive sender's evolved ability to lie are typically stable.[15] It is the payoff structure of the second-stage game that drives the evolutionary process whereby the executive sender learns what the terms have evolved to mean, learns when they might be used for gain, then exploits their evolved meanings in precisely those circumstances by lying when the opportunity presents

---

[14] See Argiento, Pemantle, Skyrms, and Volkov (2009) for a proof of this result.

[15] See the appendix for details regarding the simulations.

itself. This allows for both successful communication in the cooperative context in which the meanings of the terms initially evolved and for their occasional deceptive use by the sender.

Appendix to the example

In the learning-to-lie model the cumulative success rate of the sender and receiver on the first-stage basic signaling game is typically (0.997 of the time) better than 0.80 on simulation after one million plays. The second-stage game starts with the basic dispositions that evolved in the first-stage, then continues for another million plays. The success rate of the executive sender lying when the context presents a *safe chance* for deception is 0.896. When the context is *business as usual* the executive sender rarely lies, and the receiver, hence, nearly always does the right thing given the current state. As a result, the receiver is typically (0.968 of the time) nearly as successful as possible (just under a 0.75 cumulative success rate overall). Since both states of nature and opportunities for deception are unbiased, the probability of the context being *safe chance* is (1/2)(1/2)=1/4, so when the executive sender successfully evolves the ability, he lies about a quarter of the time in the second-stage game.

While a full analysis of model goes beyond the scope of the present paper, there are a few things worth noting. First, both the first-stage and second-stage games are relatively robust under different payoffs as long as they exhibit the same basic structure as the payoffs described above. If the payoffs in the second-stage game are changed so that successful deception pays off with one ball instead of two, for example, the agents are each just slightly more successful in their aims. Here the executive sender has a cumulative lying success rate of 0.923 and the receiver is nearly as successful as possible 0.975 of the time on simulation.

The model is somewhat more sensitive to the rate of opportunities for deception. If lying is too common in the second-stage game (significantly higher than 25%), the executive sender's attempted deceptions will eventually undermine the evolved meanings of the terms. On the other hand, if opportunities for deception are too rare, the executive sender will not learn to individuate the two natural contexts *business as usual* and *safe chance* as reliably on simple reinforcement learning. This is because suboptimal partial pooling equilibria are increasingly common under simple reinforcement learning the stronger the natural bias in the states being individuated. In such situations, reinforcement learning with punishment or forgetting, win-stay/lose randomize, or probe-and-adjust work much better to individuate the states. See Barrett and Zollman (2008) for a general discussion of this phenomena.

References:

Argiento, R., R. Pemantle, B. Skyrms and S. Volkov (2009) “Learning to Signal: Analysis of a Micro-Level Reinforcement Model” *Stochastic Processes and their Applications* 119(2): 373-390.

Barrett, J. A. and B. Skyrms (2017) "Self-Assembling Games" *British Journal for Philosophy of Science* 68:329-353.

Barrett, J. A., B. Skyrms, C. Cochran (2018) “Hierarchical Models for the Evolution of Compositional Language” manuscript.

Barrett, J. A. and K. Zollman (2008) “The Role of Forgetting in the Evolution and Learning of Language” *Journal of Experimental and Theoretical Artificial Intelligence* 21(4): 293-309.

Birch, J. (2014) "Propositional Content in Signaling Systems" *Philosophical Studies* 171:493-512.

Dretske, F. (1981) *Knowledge and the Flow of Information* Cambridge, Mass.: MIT Press.

Fallis, D. and P. Lewis (2017) “Toward a Formal Analysis of Deceptive Signaling” forthcoming in *Synthese.* DOI 10.1007/s11229-017-1536-3.

Flower, T. (2011) "Fork-Tailed Drongos use Deceptive Mimicked Alarm Calls to Steal Food" *Proceedings of the Royal Society B.* 278:1548-1555.

Godfrey-Smith, P. (2012) "Review of Brian Skyrms' *Signals*" *Mind* 120:1288-1297.

Harms. W. (2004) "Primitive Content, Translation, and the Emergence of Meaning in Animal Communication" in D. K. Oller and U. Griebel eds. *Evolution of Communication Systems: A Comparative Approach* Cambridge: MIT Press 31-48.

Huttegger, S., Skyrms, B., Smead, R. and Zollman, K. (2010) "Evolutionary Dynamics of Lewis Signaling Games: Signaling Systems vs. Partial Pooling" *Synthese* 172: 177-191.

Lewis, D. K. (1969) *Convention* Cambridge: Harvard University Press.

Lewis, S. M. and Cratsley, C. K. (2008) "Flash Signal Evolution, Mate Choice and Predation in Fireflies" *Annual Review of Entymology* 53: 293-321.

Martinez, M. (2015) “Deception in Sender-Reciever Games” *Erkenntnis* 80: 215-227.

Maynard Smith, J. and Harper, D. (2003) *Animal Signals* Oxford: Oxford University Press.

Millikan, R. G. (1984) *Language, Thought, and Other Biological Categories* Cambridge: MIT Press.

Searcy, W. A. and Nowicki, S. (2005) *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems.* Princeton: Princeton University Press.

Shea, N., Godfrey-Smith, P, and Cao, R. (2017) "Content in Simple Signaling Systems" *British Journal for Philosophy of Science*. https://doi.org/10.1093/bjps/axw036.

Skyrms, B. (2010) *Signals: Evolution Learning and Information* Oxford and New York:Oxford University Press.

Sterelny, K. (2003) *Thought in a Hostile World: The Evolution of Human Cognition* Oxford: Blackwell.

Wheeler, B. C., & Hammerschmidt, K. (2013). Proximate Factors Underpinning Receiver Responses to Deceptive False Alarm Calls in Wild Tufted Capuchin Monkeys: Is It Counterdeception? *American Journal of Primatology 75*: 715–725. http://doi.org/10.1002/ajp.22097

Wheeler, B. C. (2009) “Crying Wolf? Tufted Capuchin Monkeys Use Anti-Predator Calls to Usurp Resources from Conspecifics” *Proceedings of the Royal Society of London B: Biological Sciences* 276 (1669): 3013-3018.

Zollman K., Bergstrom C., Huttegger S, (2013) "Between cheap and costly signals: the evolution of partially honest communication" *Proceedings of the Royal Society of London Series B:* 280: