NATURALIZING NATURAL SALIENCE

JACOB VANDRUNEN AND DANIEL A. HERRMANN

Abstract. Grice, Lewis, and Skyrms proposed similar distinctions between kinds of meaning. The meaning of terms in human language, as Lewis and Skyrms had it, is 'conventional'. Skyrms presented models showing how it is possible for conventional meaning to evolve in a population without reliance on pre-existing meaning. But one might think of conventionality as coming in degrees, based on whether the evolutionary process begins with 'natural saliences'. We propose a theory of natural salience and several extensions of Skyrms's models to capture this notion. These models reveal that natural saliences can hinder, as well as help, the evolution of language.

Contents

1. Meaning and Natural Salience

The difference between 'natural' and 'non-natural' meaning, as Paul Grice had it, is the difference between the uses of 'meant' in the following sentences:

(1) The spots meant that you had measles.

(2) When the doctor said 'You have measles', she meant that you had measles.

This paper is forthcoming in the British Journal for the Philosophy of Science. Article DOI: https://doi.org/10.1086/725654.

In (2), the doctor meant something non-naturally, because she 'intended the utterance ... to produce some effect in [her] audience by means of recognition of this intention' (Grice,[1957], 385). Cases like (1) lack this feature. The spots don't intend to tell you that you have measles, nor do you recognize such an intention. The spots just indicate that you have measles, naturally.

Brian Skyrms, following David Lewis, suggested that Grice's distinction would be better put as one between 'non-conventional' and 'conventional' meaning. Grice's non-natural meaning becomes conventional meaning for Lewis, ([1969], 159). For Skyrms, all meaning is natural in the sense that it 'depends on associations arising from natural processes' ([2010], 1). The doctor and you have learned, over the course of your lives, the linguistic conventions necessary for her to communicate your diagnosis. Learning is a natural process. The spots, on the other hand, have learned nothing and follow no such conventions. They just mean measles in the same way that smoke means fire.

In his doctoral thesis on convention, Lewis used game theory to demystify the claim that language is conventional. At the time, game theory had focused on scenarios in which the interests of agents conflicted. Lewis recognized the importance of the coordination problem, in which the interests of agents align. The solution is conventional when the following three conditions are met: (i) a coordination problem has multiple solutions (equilibria), (ii) agents prefer most of all that everybody adheres to the same solution, and (iii) for that reason all agents actually adhere to the same solution (Lewis,[1969], 42). Using a class of coordination problems known as 'signalling games', Lewis argued that language fits the bill.

Although Lewis was content to take the analysis this far, Skyrms had in his sights a more difficult question: how might linguistic conventions arise without recursive reliance on prior language? This question traces its origins back through the Epicureans to Jean-Jacques Rousseau and Lewis's own doctoral advisor W. V. Quine.[1] Put another way: how is it possible for pre-linguistic agents to become (proto-)linguistic agents? A thoroughgoing naturalism about the origins of language requires an answer to this skeptical challenge, and Lewis's account does not suffice. Lewis appealed to three processes for equilibrium selection: prior agreement between agents, precedent, and salience. Prior agreement and precedent are off the table due to the persistent threat of circularity (Skyrms,[2000], 81). And where would salience come from? Charles Darwin offered one such conjecture:

> Since monkeys certainly understand much that is said to them by man, and when wild, utter signal-cries of danger to their fellows; and since fowls give distinct warnings for danger on the ground, or in the sky from hawks (both, as well as a third cry, intelligible to dogs), may not some unusually wise ape-like animal have imitated the growl of a beast of prey, and thus told his fellow-monkeys the nature of the expected danger? This would have been the first step in the formation of language. ([1875], 87)

The idea is that there was a natural salience to the terms in the simple languages of the earliest hominids. Through a tortuous process of evolutionary modification, these asymmetries in the initial conditions of linguistic development eventually birthed the meanings of terms in the languages we have today.

[1]See Rousseau's Second Discourse ([2011], 57-58) and, for example, (Quine, [1936]).

The problem, of course, is that Darwin's conjecture is likely wrong for human languages. Looking at modern human verbal communication, we find no evidence of Darwinian natural salience at work—save in the case of onomatopoeia. Skyrms, however, used Lewis's signalling games to show what Darwin and Lewis could not: how the chancy natural processes of evolution and learning can break initial symmetries and lead to the emergence of simple languages with meaningful terms. Despite Skyrms's success, it is clear that meaning in nature often does arise in the presence of asymmetries. Put this way, the question becomes, not whether we can get communication without asymmetries, but how the reintroduction of asymmetries might affect the development of language.

We are concerned in this paper with two of Skyrms's suggestions regarding natural salience. First, Skyrms suggested that incorporating natural salience into his signalling game models did not present any special difficulties ([2010], 21). This is a simplification which warrants expansion. There are many ways in which asymmetries can be introduced into signalling game models. Such asymmetries are also present in nature, giving these models additional degrees of realism. Using recent developments in the theory of self-assembling games, we can also take the analysis one step further: giving an account of how natural processes might give rise to natural saliences in the first place. Such an account bolsters the Skyrmsian position that all meaning is the result of natural processes.

Second, Skyrms suggested that the case without any natural salience is the hardest case for signalling to get off the ground in ([2010], 8). As we address the first suggestion in increasingly subtle ways, we will see that this second suggestion sometimes gets things backward. The processes which drive natural salience can also drive a wedge between agents' pragmatic success (in terms of reward) and successful communication (in terms of information transfer). This is not a refutation of Skyrms's analysis. Rather, it shows how critical symmetry-breaking adaptive processes like evolution and learning are to the successful use of language.

2. Learning to Signal

A basic $N \quad N \quad N$ Lewis signalling game has three players: nature, the sender, and the receiver. Nature picks one of N possible states. The sender observes the state of nature, and sends one of N possible signals. The receiver observes the signal, and performs one of N possible acts. Each of the possible acts corresponds to one of the states of nature. If the receiver's act matches the state of nature, both sender and receiver receive a payoff. There exist $N!$ possible pure strategy profiles in which the sender maps each state of nature to a different signal, and the receiver maps each signal to the correct act. These are called the separating equilibria, and the corresponding term language that the agents use is called a signalling system. But there are many other inefficient equilibria in the game, in which the term language the agents use contains synonyms and information bottlenecks, but in which neither the sender nor the receiver could do better by unilateral deviation from the chosen strategy. These are called the partial pooling equilibria. In a signalling game, the development of a simple term language becomes a matter of selecting an efficient equilibrium. Agents require no help from a pre-existing language if they can select these equilibria in some other way.

Lewis told a qualitative story about how random chance might produce coordination in high-rationality agents ([1969], 39). Agents coordinate by happenstance,

Figure 1. A 2 2 2 Lewis-Skyrms signalling game with act-based reinforcement learning.

and the memory of that initial serendipitous coordination event generates mutual dispositions to act similarly in the future. Skyrms gave this story quantitative rigor, showing that successful coordination requires only minimal assumptions about agents' rationality. Speci cally, their behaviours are plastic in the way described by Thorndike's law of e ect: successful actions end up being performed more often in the future (Thorndike, [1911]). This process of behaviour modi cation is called reinforcement learning. Skyrms showed that organisms well-modeled by these conditions are capable of learning a simple term language to describe their environment. The insight is that linguistic conventions can be established *ex nihilo* even among agents with minimal cognitive resources.

A common learning process used in Skyrms signalling game models is `simple reinforcement learning.[2] We can characterize it with balls and urns. The sender has N urns, each corresponding to a state of nature. The sender draws from the corresponding urn after observing nature. The urn has balls of N di erent colours, each corresponding to a possible signal. The sender sends the signal corresponding to the ball drawn. The receiver in turn has N urns corresponding to the possible signals, each with balls of N colours corresponding to the possible acts. When the sender and receiver are jointly successful, they return the balls drawn to their respective urns, and add an additional ball of the same colour. When they are not successful, they return the balls to the urns without adding any new balls. Figure 1 illustrates an example setup for the case of $N = 2$.

Under simple reinforcement learning, Hu et al. ([2011]) proved that every signalling system has a positive probability of being attained in the limit. Argiento et al. ([2009]) proved for the special case of $N = 2$ that there is long-run convergence to a signalling system with probability one. But for the more general case of $N > 2$, as far as we know, it is not guaranteed that the players will arrive at a signalling system in the long run. Je rey Barrett ([2006]) showed by simulation that in the medium run players in signalling games with $N > 2$ sometimes end up in partial pooling equilibria, and that this e ect increases with larger N and when nature is biased. Even so, Barrett showed that agents often select relatively e cient equilibria, and that more sophisticated learning processes make it more likely that agents will end up in signalling systems.

2.1. What Natural Salience Is Not. Now that we have the toolkit set up, we must frame our discussion of natural salience against the backdrop of other

[2]This learning model is motivated both psychologically and by its formal relation to the evolutionary replicator dynamics. See (Skyrms,[2010], ch. 7-8) for background on the model and its application to evolutionary signalling games. Note that this is an act-based implementation of the dynamic game: individual acts|rather than whole strategies|are being reinforced.

Skyrmsian investigations of salience. Recently, there have been two such projects: Travis LaCroix's `salience games' and several authors' work on the theory of self-assembling games. These projects fill important gaps in Skyrms's account and give us powerful modeling resources to draw from. But they do not touch the issues we are concerned with in the present paper.

LaCroix ([2020b]) tested the hypothesis that introducing salience into a Skyrms-style signalling game decreases the probability of agents ending up in partial pooling equilibria. He tested this by introducing a `salience parameter' which allows for a smooth transition between Lewis and Skyrms. The idea is that most of the time agents behave like simple reinforcement learners. But with some probability , the sender will instead send a signal corresponding to the most-reinforced ball in the relevant urn, rather than drawing from the urn at random. Ditto for the receiver, mutatis mutandis. This amplifies the success of previously-successful actions beyond what simple reinforcement learning is capable of doing. LaCroix showed by simulation that agents in his salience games avoid partial pooling equilibria significantly more often than vanilla reinforcement learners do, and that they reach signalling systems significantly more quickly.

In light of LaCroix's project, we come to understand salience in at least two ways. The first is what Robin Cubitt and Robert Sugden called the `salience of precedent' in a reconstruction of Lewis (Cubitt and Sugden,[2003], 201). The second is Darwin's natural salience. LaCroix salience games generalize the Lewis-Skyrms signalling game by incorporating the salience of precedent above and beyond that captured by simple reinforcement learning. The addition of LaCroix's salience parameter allows agents in Skyrmsian models to recognize this salience with different degrees of reliability. But in this paper we are concerned with saliences of the second kind: natural saliences which break initial symmetries without precedent.

Barrett ([2021]) discussed a further kind of `salience': the salience which determines what players pay attention to when choosing their acts in the first place. Daniel Herrmann and Jacob VanDrunen went on to give a generalized description of the problem:

> Suppose [the sender] signals by waving a red flag. What is the signal here? Is it the colour? The fact that it is a flag? The pattern in which she waves it? Where she stands when she is waving it? ([2022])

Barrett, Herrmann, and VanDrunen were working within the nascent framework of self-assembling games, introduced in (Barrett and Skyrms, [2017]). The fundamental question in the theory of self-assembling games is: how might signalling games come to be played in the first place? That is, how might a sender learn to become a sender, and a receiver a receiver? Part of this involves the receiver learning to recgonize that the sender's action can be interpreted as a signal about nature.

In the simplest case, Herrmann and VanDrunen's `attention game' model works as follows. The sender is just a vanilla sender as in a basic $N \times N$ signalling game. The receiver, however, has an urn which controls what she attends to when conditioning her action. One of the kinds of balls in this attention urn corresponds to attending to the sender's act. The other kinds correspond to attending to other possible things she can attend to. These are called `observation' or `signalling channels'. The receiver also has a set of action urns corresponding to these channels.

Play on the receiver's end proceeds as follows. She first draws a ball from the attention urn. This determines the channel she pays attention to. Finally, she draws from the action urn corresponding to the value of that channel, and performs the corresponding act. Both her attention urn and action urn are reinforced.

Herrmann and VanDrunen showed that the receiver can learn to attend to the sender's act while the sender and receiver simultaneously learn a signalling system. In attention games, the saliences of agents|in the sense of what they pay attention to in the world|co-evolve with their signalling dispositions. [3]

Attention games generalize signalling games in a critical way. A basic signalling game has the available signals baked in: a set of phenomena are naturally salient to the receiver as possible signals to be acted on. But in attention games, the receiver starts without the natural salience of one set of signals. Here we have a more thoroughgoing attempt at eliminating the need for a philosophical appeal to natural saliences.[4]

These considerations locate the present account within the landscape of salience. Here we investigate natural saliences, rather than the saliences LaCroix was concerned with. Furthermore, we are concerned with the saliences of particular signals within a signalling channel, rather than the saliences of particular signalling channels (in contrast to Barrett, Herrmann, and VanDrunen's original models). Despite the difference in the type of salience modeled, our account builds on the mechanics of previous models, and makes contact with previous questions. In §4.1, for example, the model is a direct extension of the basic attention game. Likewise, in the spirit of LaCroix's analysis, a principal concern will be whether or not the presence of natural salience helps agents escape from partial pooling equilibria.

Now that we understand our situation with respect to prior work, we can turn to a discussion of natural salience in more detail. We begin with a survey of the empirical data surrounding natural salience in nature.

3. Varieties of Salient Experience

Darwin's conjecture about natural salience extended beyond the evidence he cited. His evidence concerned the more general case of animals picking out particular kinds of predators with their alarm calls. The vanilla Lewis-Skyrms framework deals with this case well. At issue in his conjecture about the language of early hominids, however, is more than their ability to pick out natural kinds by distinct vocalizations. It is whether the initial vocalizations they used were somehow well-suited to denote the kinds they picked out. We are concerned in this section with what forms this `well-suitedness' might take in nature. We suggest that the phenomenon of natural salience might be divided into two categories: positive and negative natural salience.[5] We discuss each in turn.

---

[3] Attention games thus answer another issue that Cubitt and Sugden raised: the need for a `theory of the co-evolution of conventions and of concepts of salience' ([2003], 202).

[4] This need might be reduced in other ways. For example, Alexander et al. ([2012]) provided a set of signalling game models in which agents invent new signals within a fixed channel. This contrasts with the attention game models, in which agents learn which channel to communicate across in the first place.

[5] The designations `positve' and `negative' are themselves conventional. They may have carried natural meanings at the genesis of the present project, but the authors are no longer able to articulate them.

Positive natural salience occurs when a signal is well-suited to its use in virtue of its prior association with the phenomenon it is used to denote. Before agents begin communicating, one or more of the agents has experienced some stimulus in conjunction with a phenomenon of interest. Later, when the agents want to establish a communication system, a signal imitating that stimulus might then end up being used as a term denoting the already-associated phenomenon in the language the agents develop.

Darwin had positive natural salience in mind when he posed his wise-ape conjecture. In nature, the most common way in which communication based on positive natural saliences comes about is through the process of ritualization.[6] For example, some caterpillars are able to decide territorial disputes by ritualized displaying in which a territory's `owner' will produce vibrations with its rear segment and mandibles. Scott et al. ([2010]) nd evidence from comparative morphology that these vibratory displays originated as phenomena associated with violent territorial defense. Over time, these vibrations become recognizable as signals of territorial ownership. This simple communication system is incentivized by the mutual bene t derived from peaceful con ict resolution.

Consider again our de nition of positive natural salience: a signal being well-suited to its use in virtue of its prior association with the phenomenon it is used to denote. In the caterpillar example, the vibratory stimulus has become a signal by which the owner communicates its territorial claim to the intruder. Even if the language of caterpillars has conventional features, the natural salience of particular vibrations existed before, and presumably informed, the caterpillars' conventional system of communication.

Darwin suggested the case of vocal alarm calls which imitate a predator in order to communicate the predator's presence. The presence of this phenomenon in nature is more dubious, but some evidence for it among birds exists. The Sri Lanka Magpie will sometimes imitate the call of one of its predators, the Besra Sparrowhawk. Ratnayake et al. ([2009]) observed this happening in conjunction with increased magpie mobbing of nearby hawks. Other evidence is scarce. In an extensive literature review published shortly before the magpie study, Kelley et al. ([2008]) considered a variety of attempts to explain vocal mimicry in birds. Due to a variety of hypotheses on o er and the lack of decisive evidence discriminating among them, they concluded that `we are no closer [than we were in 1982] to determining even a single function for vocal mimicry [in birds]' ([2008], 526).

Although common in birds, vocal mimicry is absent in primates (Planer and Sterelny,[2021], 158). In mild support of Darwin's conjecture, vocal mimicry does form a staple of communication in some contemporary foraging societies (Lewis, [2009]). Planer and Sterelny suggest that the utility of vocal mimicry may have brought ne-grained vocal control under positive selection pressure in hominids ([2021], 78). But the models we will consider in §4 demonstrate that developing imitative alarm calls is a complex process. And, as it turns out, this process can pull apart the formation of linguistic systems from the formation of pragmatically successful behaviour. We are thus left with little reason to expect Darwinian mimicry to be an accurate model for the development of human language. So much for positive natural salience.

[6] For classic treatments of ritualization in zoology, see (Tinbergen; Huxley; Lorenz, [1952; 1966; 1966]).

Negative natural salienceoccurs when a particular signalling convention is well-suited in virtue of the context in which it is deployed. There are at least two ways in which this can occur. In the rst, the use of particular signals is di erentially rewarded in particular states of the world. In the second, the use of signalling conventions which prioritize certain states of the world is di erentially rewarded.

Kavanagh ([1980]) reports a remarkable example of what we call negative natural salience in the Tantalus Monkey (a species of vervet). Some vervets in Cameroon live on the savannah. These are prey to wild dogs. Others live in forests, where they are prey to human farmers who hunt the pests with domesticated dogs. On the savannah, vervets use a loud alarm call to signal the presence of dogs. But in the forests, vervets use a quiet call to signal the same predator. This allows the vervets to make an inconspicuous escape from the more sophisticated hunting weapons employed by the dogs' human companions.

Returning to our de nition of negative natural salience: the forested vervet's signal for `dog' has become well-suited in virtue of an advantage its use a ords in the presence of the phenomenon it is used to denote (dog [with human]). Negative natural salience arises because di erent signals incur unequal costs.

In the above examples, certain signalling systems have a natural salience because their particular terms are more e cient at signalling about states. One can also imagine a case in which certainstates are more important to talk about, full stop. Take, for instance, a hypothetical prosimian species whose members are primarily preyed upon by lions. They might have many things they could use their language to talk about. But suppose that they have limited communicative resources, so they must choose their words wisely. It makes some sense to think that a likely language for them would partition the world into `lion' and `everything else', provided that their use of language is motivated by pragmatic, evolutionary success. This is another case of negative natural salience: one in which the signalling convention as a whole, including its partitioning of the world, is salient.

We will discuss models of negative natural salience later in§5. But rst we turn to positive natural salience.

4. Positive Natural Salience

Positive natural salience removes the initial symmetry in a signalling game by altering the initial propensities of the agents. Skyrms writes:

> The [Darwinian] scenario of some small initial natural salience ampli ed by evolutionary feedback may well be the correct one for many evolutionary histories. [. . .] It can be represented in signalling games simply by moving the initial probabilities o exact symmetry|in a given state the sender is initially more likely to send one particular signal than others, and a receiver is more likely to react to that signal in the appropriate way. ([2010], 21-22)

One way this kind of symmetry-breaking might be accomplished in a vanilla Lewis-Skyrms signalling game is by adding extra initial balls of certain colours to certain urns in order to jump-start the learning process. Contrastingly, the basic model presented here shows how the initial symmetry might be broken via the dynamics of pre-play learning, eliminating the ad hoc stipulation of favourable initial conditions. Here it is important to distinguish between the agents' history prior to playing the signalling game, and the agents' history of play in the signalling

game. The goal of the basic model is to show how the receiver's history prior to her becoming the receivermight a ect her propensities when she initially begins playing the signalling game with the sender. This generalizes the stipulation of initial asymmetries, embedding the saliences of an agent within the context of a broader learning process which extends beyond the signalling game itself. The model subsequently presented in §4.1 adds further degrees of freedom in an attempt to provide an even more general account of this process.

For the basic model, which we will call Model A, consider a modi ed $N \times N \times N$ Lewis-Skyrms signalling game with $N = 16$ to ensure a large number of viable partial pooling equilibria. Sender and receiver both begin with one ball of each type in their urns, but the sender does not play in the early rounds of the game. For an initial sequence of K rounds, the signalling channel is instead directly correlated with nature in the following way. With probability p, the signal sent maps perfectly onto the state of nature according to one of the $N!$ possible signalling systems (pre-selected at the outset of the game). With probability $1 - p$, the signal sent is chosen uniformly at random. This can be thought of as the receiver imperfectly observing the actionable state of nature via some correlated cue. It falls into the category which Barrett and Skyrms ([2017]) call a cue-reading game, albeit with only partially-informative cues.[7] During the initial rounds, the receiver learns by simple reinforcement. For all rounds after the rst K, the signalling channel is manipulated by the sender's play instead. At that point, the sender begins to learn by simple reinforcement as well. The results reported below are based on 1000 simulations each of the following experimental conditions: $p \in [0, 1]$ with increments of 0.1 and $K \in \{10^2, 5 \cdot 10^2, 10^3, 5 \cdot 10^3, 10^4\}$, with 10^7 regular rounds following the initial rounds.

Appendix A contains a detailed description of the results. The basic nding is that both increasing the correlation p of the initial signalling channel and increasing the number of initial rounds K leads to the attainment of more robust signalling conventions. Agents' accuracies increase as both p and K increase. Senders will also more reliably learn to use parts of the pre-selected signalling system that the channel was initially associated with. That is, if one state was associated with signal during the initial rounds, a sender will subsequently be more likely to use to represent that state, as p and K increase. But these e ects are more sensitive to increases in K than they are to increases in p. Even a slightly-correlated channel can have a large impact on the nal signalling convention if the receiver has been exposed to it for a long period of time.

Model A provides an account of Darwin's positive natural salience within the framework of Skyrms signalling games. The objective initial correlation of the signalling channel produces a subjective salience of particular acts for particular observed signals on the receiver's part. The sender, in turn, learns to mimic the signals sent during the initial rounds, playing what Barrett and Skyrms ([2017]) call a sensory-manipulation gamegiven the receiver's initially-learned dispositions. This salience exempli es two properties. First, it is dynamically acquired from initially-symmetric conditions: all that was required was for there to exist a conditionally correlated observation channel which the receiver happens to be attuned to. Second, as a basis for Lewisian common knowledge, it is weak|perhaps about the weakest

[7] More speci cally, the mutual information between the observation channel f_1 and the nature random variable in the initial rounds is $I(f_1;) = [(N 1)p + 1] \log[(N 1)p + 1]$.

Figure 2. Model A/1 with $N = 4$.

one could go. Lewis writes that salience in general is a weak basis for common knowledge because `the salience of an equilibrium is not a very strong indication that agents will tend to choose it' (Lewis,[1969], 57). In this case, in fact, the salient equilibrium is not even salient to both players. The Skyrmsian adaptive dynamics is still doing the heavy lifting for establishing the signalling convention, but positive natural salience is getting it o the ground in a particular direction.

The largest remaining gap in the story is how the agents coordinate on the conditionally correlated channel in the rst place. How does the receiver become attuned to the relevant feature of the world, and how does the sender come to manipulate that feature? In the next section, we show how this can come about by means of an adaptive process as well. This will also shed light on how the weak salience we are studying can produce failures of coordination as well as successes.

4.1. Extended Attention Games. The next model is a modi ed Herrmann-VanDrunen attention game. In the standard attention game, features of the world which serve as potential signalling channels are correlated with the sender's action. In the modi ed attention game presented here, signalling channels are instead correlated with nature.[8] There is also an additional process by which a sender can choose which channel to manipulate. Both sender and receiver must now learn not only to signal, but to play the signalling game with each other in the rst place. This setup allows us to examine how a sender and receiver might learn to coordinate on a particular signalling channel based on its conditional correlation with nature.

We will now consider the model, which we will denote Model A/1, in detail. The game has N states of nature, N corresponding acts, and 2 possible signalling channels (features) f_0 and f_1, both of which can take on N di erent states. The dispositions of the players may once again be conceptualized as urns. In addition to two sets of N urns mapping states of nature to signals (one set for each feature),

[8]That is, in the original attention game, features are correlated with nature only through their correlation with the sender's action, which is in turn correlated with nature. But here we have the opposite: the channel is correlated directly with nature, and thus also possibly with the sender's actions as mediated by nature.

the sender also has a *manipulation urn* with balls of 2 colours corresponding to the different signalling channels he could choose to manipulate. On a round of the game, the sender draws a ball from his manipulation urn, then draws a ball from his signalling urn corresponding to the state of nature and the chosen signalling channel, and sends the corresponding signal.

The signalling channel which the sender does not manipulate takes on a value based on the following algorithm for conditional correlation. If it is f_0, it takes on any value uniformly at random—that is, f_0 is perfectly uncorrelated with nature when the sender is not manipulating it. If it is f_1, however, it takes on a value using the same correlation procedure described for the initial rounds of Model A: with probability p it takes on a value corresponding to a fixed bijective map from states of nature onto signals, and with probability $1 - p$ it takes on a value uniformly at random.

The receiver likewise has two sets of N action urns corresponding to the two possible signalling channels, and an *attention urn* with which she selects which channel to condition her act on. Sender and receiver reinforce all urns at the end of the round with the simple reinforcement dynamics. The complete model is illustrated in Figure 2. Once again, the reported results are based on 1000 simulations for 10^7 rounds of play, but this time there are no initial rounds. The experimental conditions tested are $p \in [0, 1]$ with increments of 0.1 and $N \in \{2, 4, 8, 16\}$.

Appendix B contains a detailed description of the results, along with helpful illustrations. Here, the findings are significantly less intuitive. As p increases from 0 to middling values, the mean accuracy of agents *decreases*, before increasing again to above the baseline as p approaches 1 (perfect correlation of f_1 with nature). This effect, as well as the effects described below, become more dramatic as N (number of states/signals/acts) increases.

Further examination sheds light on this counterintuitive result. First of all, as p increases, receivers are more likely to pay attention to f_1. So far, this makes good sense. But as p increases, senders become *less* likely to manipulate f_1, after a slight initial increase in the likelihood for low but non-zero values of p. When we focus in on only those senders and receivers who successfully coordinated on f_1, we see two additional trends. For middling values of p, the senders in this group (as in Model A) are more likely to be using parts of the signalling system that f_1 was conditionally correlated with. The mean accuracy of the sender/receiver pairs in this group is likewise higher than the mean accuracy of those who didn't coordinate on f_1. But both of these quantities (alignment with the naturally salient signalling system and relative accuracy) peak and then decrease as p approaches 1.

By the time $p = 1$, agents who successfully coordinated on the conditionally correlated channel f_1 are doing worse on average than agents who didn't coordinate on f_1. This is the most counterintuitive result of all, but understanding it is the key to the whole thing. The group of agents who didn't coordinate on f_1 consists of both those agents who coordinated on f_0 instead, and those agents who didn't coordinate at all. In the group of agents who didn't coordinate, senders learn to manipulate f_0, and receivers learn to manipulate f_1. If $p = 1$, receivers in this group will learn to do the correct act with perfect accuracy. If the sender manipulates f_1 and overrides the default correlation with nature, the receiver's accuracy can only go down. Because the sender is reinforced not on success in communicating with the receiver but on the the receiver's pragmatic success, the (would-be) sender learns

not to talk to the (would-be) receiver. Again, the interested reader may refer to Appendix B for a more detailed and visual presentation of the results.

The basic conclusion here is that the presence of a signalling channel which has a small conditional correlation with nature can improve both the chances of coordination and the effectiveness of signalling when players coordinate on it. This supports our findings from Model A. But this basic conclusion is tempered by the further result that as conditional correlation increases, the chances of miscoordination also increase. For middling values of p, this is to the detriment of the success of the players as a whole—the receiver is attracted to the conditionally correlated channel, but without sender intervention is not capable of doing better than the conditional correlation allows. But when p approaches 1, the receiver is capable of near-perfectly-successful action even without sender intervention.

That the receiver will do better without sender intervention is obviously the case at $p = 1{:}0$, but is less obviously the case even at $p = 0{:}9$. When $p = 0{:}9$ and $N = 16$, the greatest possible accuracy (without sender intervention) is $9/10 + (1/10)(1/16) \approx 0{:}906$. This is slightly less than the average accuracy of $0{:}908$ attained in the regular $16 \times 16 \times 16$ Lewis-Skyrms signalling game (Model A, $K = 0$). But initially—before the sender has learned any signalling conventions—the receiver will do better without sender intervention as long as $p > 0$. It is this factor which allows miscoordination to take hold in the learning process.

There are two kinds of symmetry-breaking that must happen in attention games: within-channel coordination of which signals mean what, and between-channel coordination of which partition of the world to treat as a signal in the first place. Conditional correlation can lead to a natural salience that affects both processes. It can facilitate within-channel coordination, as seen both in Model A and in Model A/1 for middling values of p. It can also facilitate between-channel coordination, as seen for small values of p in Model A/1. But it can also impede the process of between-channel coordination, as illustrated in Model A/1 with middling-to-high values of p. The sender is rewarded based on whether the receiver is successful, and not whether the receiver is successful *by attending to the channel the sender is manipulating*. So, there is an initial pragmatic cost to the sender's choice of breaking the conditional correlation if the receiver could be doing better than chance without sender intervention.

Skyrms assumes that the case without something like positive natural salience is the worst case for the emergence of coordinated signalling conventions:

> In some cases there may well be natural salience, in which case the amplification of pre-existing inclinations into a full fledged signalling system is that much easier. ([2010], 21)

We have seen that Skyrms is correct for the case of within-channel coordination, but Model A/1 shows that there is an even worse case for between-channel coordination. This is when there are possible signals which are *too* well-suited to convey their content. In such cases, it is not worth it for the sender to signal at all.

Another gap remains in the model if we want it to capture Darwin's notion of vocal mimcry in alarm calls. Suppose a predator's presence or absence is the state of nature, and the predator's presence is correlated with a particular sound. But while the predator doesn't always make the sound (imperfect correlation), the (proto-)sender can imitate the sound to alert his partner. On the surface, this bears many similarities to the situation in Model A/1. But consider: if the predator has

growled, the sender cannot make the predator run-growl. Thus far, our signalling game models have assumed that the sender has perfect control of whether or not the receiver does or doesn't hear a growl. But this is not the case in general. Usually, an asymmetry in the possibility of intervening on a particular signalling channel exists. The next model fills this gap.

4.2. Asymmetrical Intervention. To capture the case of an asymmetrical possibility of intervention, we can consider the following modifications to the game in Model A/1. Call this Model A/2. The state of nature is that a predator is either present or absent (50/50 chance), thus we only consider the case where $N = 2$. There are two possible vocalizations which the sender can make, and which the receiver can hear. The sender, upon observing whether or not there is a predator, chooses among the acts `make sound ', `make sound ', and `do nothing'. Making sounds and can be thought of as manipulating channels f_0 and f_1 respectively. Independently of the sender's choice, with probability p, the predator (if present) will make sound . Thus, f_1 is correlated with nature. The sender's possibility of manipulating f_1 is limited: if the predator did not make sound , the sender's decision to make sound will change the value of f_1. (Assume that by default things are silent.) If the predator did make the sound, however, nothing the sender can do will change the value of f_1 (although the sender might redundantly choose to make sound). The receiver, as before, chooses between listening for and listening for .

The behaviour of agents|in particular the sender|in Model A/2 warrants further investigation. Appendix C contains a detailed discussion of the results. Receivers always learn to be very successful. But what we find on the whole is that the same pattern is present in Model A/2 as was present in Model A/1. If the reader had hypothesized that an asymmetry in the possibility of intervention would close the gap between the raw success of the agents playing the game and the coordination of the agents on the conditionally correlated signalling channel, the reader would have been mistaken[9]. By all appearances, this phenomenon is ubiquitous in attention games with correlated signalling channels. These results serve to highlight a critical but thus-far underappreciated point: in realistic signalling scenarios, the pragmatic success of agents in the short- and medium-run often conflicts with the long-run success of agents in communicating.

5. Negative Natural Salience

Negative natural salience can develop when a signal is well-suited in virtue of an advantage its use affords in the presence of the phenomenon it is used to denote. The simplest way to capture this in a signalling game model is by stipulating that while $N! - 1$ of the possible signalling systems pay off 1 on average, the remaining signalling system pays off $r > 1$ on average. This would be the signalling system which makes use of the negative natural salience. Figure 3 illustrates this in extensive form. It is a setup similar to an example coordination game given by Lewis, ([1969], 10).[10]

[9] This is in fact what the authors speculated early on.

[10] It is important to note for the case of negative natural salience that the relative advantage of one signalling system over another does not necessarily violate Lewis's definition of convention. The primary criterion for a convention is that it is not a unique solution to the coordination problem at hand. An extreme case of negative natural salience might collapse this.

Figure 3. Top: vanilla Lewis signalling game with $N = 2$ in extensive form. Bottom: Model B with $N = 2$, well-suited to communicate A, and well-suited to communicate B.

Will negative natural salience help agents avoid partial pooling equilibria? In the context of simple reinforcement learning, we have previous evidence that the answer is Yes. This evidence comes from a model due to Skyrms, ([2010], 96-97). Skyrms's model features a related setup in which the initial weights (number of balls in agents' urns) are shifted. Skyrms found that as the initial weights of the simple reinforcement learning process decrease in magnitude, agents are less likely to end up in partial pooling equilibria. Decreasing the magnitude of the initial weights increases the magnitude of the reinforcements relative to the initial weights. This is similar to what the process giving rise to negative natural salience does: increase the magnitude of rewards for the success of certain acts, relative both to the initial weights and to the rewards for the success of other acts.

Our model operationalizes negative natural salience more directly. The model—Model B—is exactly like a vanilla Lewis-Skyrms signalling game, with the following modification. If the receiver is successful, the agents both receive a payoff of 1 (one ball added to each urn), as in a vanilla game. If, however, the agents were successful *and* the signal the sender used corresponded to a pre-selected bijection of states to signals, the agents both receive a payoff of r (see Figure 3). That bijection corresponds to the signalling system which has negative natural salience for the agents. As with Model A, we consider the case of $N = 16$, and test the effect of $r \in \{10^0, 10^1, 10^2, 10^3\}$ with simulations (note that $r = 10^0$ is just the vanilla $16 \times 16 \times 16$ game). Appendix D describes the results in detail. At $r = 10$, the agents on average learn to use about one-half of the signals in the negatively naturally salient system, and display a significant increase in mean success when compared to their success in the vanilla game. But this effect does not significantly

increase asr increases by further orders of magnitude. With a small amount of negative natural salience, agents thus typically learn to communicate using terms both with and without good reason for their use. Getting agents to a language with only terms that have good reason for their use is signi cantly harder.

5.1. Information Transfer. Allowing disparate payo s for success using di erent signals opens the door to other expansions of the Skyrmsian model. The last scenario we will consider features agents who have di erent payo s for success in di erent states of nature. This models a case in which success is critical in certain states of nature, and less important in others. For example, if a species is primarily preyed upon by lions, a language which prioritizes clear communication about lions might be selected over languages which do not prioritize talking about any particular kind of thing at all. This might come at the cost of decreasing the amount of information transferred on average.[11] Of the scenarios examined thus far, this is the most direct way in which communicative success comes into con ict with pragmatic success.

LaCroix ([2020a]) studied information transfer in signalling games with information bottlenecks. Consider a 10 2 10 signalling game|that is, one with 10 equiprobable states of nature but only 2 possible signals for the sender and receiver to communicate with. There are many possible term languages which maximize the agents' pragmatic success. If the language divides up the world so that one term picks out one state, and the other term picks out the remaining nine, this yields a mean accuracy of 0.2. The same mean accuracy is attained with a language that divides up the world into two sets of ve states. Although any language which makes use of both terms can achieve the maximum possible accuracy, only the language which divides up the world into two sets of ve maximizes information transfer.[12] LaCroix showed that agents are more likely to learn a language with high information transfer, compared to a baseline in which agents randomly select a pragmatically-successful language. There are more ways to divide the world up into two sets of ve than into a set of one and a set of nine. So, an agent who selects a pragmatically successful langage at random will already be more likely to choose a language with higher information transfer. But for a reason that is yet unclear, agents are even more likely to prioritize information transfer in simulations of the game than would be predicted by this analytic baseline.

Now consider a new modi cation of this setup. Nine of the ten states yield a baseline payo of 1 on success. One of the ten, however, yields a larger payo of $r > 1$ on success. Unlike in the previous section, the high payo no longer depends on which signal was sent, but simply on whether the receiver performs the correct action in one particular state. This is negative natural salience of a di erent kind. What becomes salient is now not the language which signals particular states with particular terms, but the languages which prioritize success in the one particular state by reserving one of the two available terms for communicating about it. Languages which prioritize information transfer are in this case less salient than languages which prioritize pragmatic success as de ned by expected payo s. Appendix E details the results from 1000 simulations of this model (Model C), with $r \in \{1, 2, 10\}$ and 10^6 rounds of play per simulation. With a small increase in r

[11] Throughout this section, by `information transfer', we will mean the mutual information between a state of the world and the signal sent by the sender, where the probabilities are given by the sender's dispositions and the biases of nature. See Appendix E.

[12] See Appendix E for details on how information transfer is calculated in signalling games.

($r = 2$), we already observe a significant deviation from LaCroix's results in favour of more pragmatically successful languages. With $r = 10$, LaCroix's results are inverted: most agents learn the languages which transmit the lowest amounts of information. To be clear, this dramatic result is—well, dramatic—because it happens after a simple modification: tweaking the magnitude of certain rewards. The observed tendency of signalling agents to maximize information transfer only holds in the knife's-edge case in which rewards are perfectly balanced.[13]

6. Conclusion

In this paper we showed how natural processes might give rise to languages that have terms which are not purely conventional. This required a theory of natural salience, motivated by lacunae in the basic signalling game models of Lewis and Skyrms (§1 and §2) and the—sometimes conjectured, sometimes observed—occurrence of such phenomena in nature (§3). We found that a variety of natural processes can create these saliences, and that the saliences can be helpful.

But the knife cuts both ways: while natural salience typically increases agents' pragmatic success, in a wide range of situations it also decreases agents' communicative success. This happens in two ways. First, natural salience can lead agents to coordinate on a language which does not maximize information transfer (§5.1). Second, it can lead agents to miscoordinate such that they do not establish a reliable communication channel at all (§4.1 and §4.2). This second way provides one possible explanation for the empirical failure of Darwin's hypothesis. Once we understand that high natural salience frequently causes the process that establishes linguistic conventions to fail, our expectation should be that whenever we observe a linguistic convention, it has arisen in a context of low (or no) natural salience.[14]

Pragmatic success can come apart from the development of communication because, in the short run, the presence of natural salience disincentivizes the sender from destroying the information that the receiver receives from nature. But this information is imperfect.[15] Furthermore, since the sender observes nature directly, in principle the agents could coordinate with a signalling system which would yield them maximal pragmatic success. Thus, the more subtle story is not that natural salience pulls apart pragmatic success and communication. Rather, it pulls apart short-term pragmatic success from the in-principle pragmatic success that could arise from a more sophisticated linguistic system.

These results are only as good as the models for which we can demonstrate them. Many extensions suggest themselves. The model of Darwin's conjecture might be further refined: for example, the base rate of a predator's presence will typically be lower than one-half. It is also typically more important to avoid a real predator

[13] Tucker et al. ([2022]) suggested a three-way trade-off in the evolution of simple term languages. The first, `complexity', is what we have been calling information transfer (mutual information). The second, `utility', is the expected payoff. The third, `informativeness', tracks task-agnostic success. It can be thought of as the expected payoff assuming that all success is rewarded equally. In the games LaCroix investigated, informativeness and utility come apart from complexity. In the game studied here, all three values come apart.

[14] Of course, to make this story precise, we would need to think about the base rates of both natural salience and linguistic conventions. Really, our model gives us insight into the likelihood of linguistic conventions given different levels of natural salience.

[15] Except for the edge case of $p = 1$, which corresponds to the context in which the receiver gets perfect access to nature.

than to avoid falling prey to a false alarm. Additionally, if a receiver is capable of paying attention to multiple channels, the sender's signal on the uncorrelated channel might become useful in spite of its redundancy.[16] Likewise for the model of negative natural salience: if naturally salient signals always bene t their users (regardless of the receiver's successful action), agents might be even more inclined to use them. Nevertheless, the models presented here instruct us well. Previously, we thought that natural salience provided a shortcut in the evolution of language. Our models witness this possibility. But they also teach us that, in many cases, natural salience hinders the development of linguistic conventions in favour of short-term pragmatic success.

Appendix A. Model A Results

Figure 4 shows the results for Model A. The most relevant quantity to consider is the `alignment with nature'. This tracks how closely the sender's signalling dispositions align with the bijection that is used to determine the value of the channel during the initial rounds (call this the `natural bijection'):

$$(3)\quad \text{Alignment} = \frac{\#\text{ of sender urns with modes corresponding to initial bijection}}{N}.$$

The baseline for this quantity when $K = 0$ is $0{:}062$, which is approximately $1{=}16$.[17] The quantity is sensitive both to increasing p and increasing K, but is more sensitive to the latter. With $p = 0{:}2; K = 10^4$, the mean alignment is 0978, which is signi cantly higher than $15{=}16$ $0{:}938$. With $p = 1{:}0; K = 10^2$, the mean alignment is only $0{:}116$, which is still less than $2{=}16 = 0{:}125$. With $p = 1{:}0; K = 5$ 10^2, however, this quantity jumps to $0{:}536$, which is around $8{5}{=}16$ $0{:}531$. We conclude from this that senders are indeed more likely to send signals corresponding to the natural bijection when the receiver has had the opportunity to learn using that bijection.

The other relevant quantity is the cumulative accuracy which players attain. The baseline at $K = 0$ is $0{:}908$. Once again, this quantity is more sensitive to an increase in K than to an increase in p. With $K = 10^2$, the players on average never do better than the baseline, even when $p = 1{:}0$. With $p = 1{:}0; K = 5$ 10^2, this jumps slightly to $0{:}926$, and it again jumps to 0984 when K is increased to 10^3. By comparison, when $K = 5$ 10^3, the mean cumulative accuracy reaches $0{:}991$ when $p = 0{:}3$, and it reaches 0993 when $p = 0{:}2; K = 10^4$. But with large K ($K > 10^3$) we observe another e ect when there is no correlation of the signalling channel in the initial rounds ($p = 0{:}0$). With $K = 5$ 10^3 the mean cumulative accuracy is below the baseline at 0902. With $K = 10^4$, it is also below the baseline at 0904. So, with a large number of initial plays, a signalling channel which is perfectly uncorrelated with nature can be more of a help than a hindrance, but with only a small correlation, it becomes a signi cant help to successful coordination on a signalling system. On the other hand, with a small number of initial plays, the initial correlation (even when it is high) has much less of an e ect overall.

[16] Such is the case in some of the models described in (Barrett, [2021]) and (Barrett and VanDrunen, [2022]). This would be particularly easy to do for Model A/2, in which the receiver could be wired to act on any of the four possible signal-observations: only, only, and , or total silence. But the modi cation would remove the attention-learning process from the receiver's end, and thus collapse the distinction between communicative and pragmatic success.

[17] Because in this case there are no initial rounds, the value of p is irrelevant.

Appendix B. Model A/1 Results

Understanding what is going on in Model A/1 is a more subtle undertaking. Consider first the results shown in Figure 5 which parallel the results from Model A. Here, because senders have two sets of signalling dispositions—one for each channel—and only one channel is conditionally correlated with nature, the alignment with nature cannot be calculated absolutely as in Model A. So, we begin by examining the cumulative accuracy. We see that accuracy decreases for middling values of p to a nadir at around $p = 0.6$ (for $N \in \{2, 16\}$) or $p = 0.7$ (for $N \in \{4, 8\}$), before rising to levels significantly above the baseline when $p = 1.0$. Unlike in Model A, accuracy is not an approximately monotonic function of p, but follows a much different pattern. What is the explanation of this?

We begin to get an idea of what's going on by examining the behaviour of senders and receivers individually. Figure 6 shows the proportion of senders and receivers who are more likely to manipulate/attend to the conditionally correlated feature f_1. As expected, the receivers display the clear trend of being more likely to attend to f_1 as p increases (the increase is monotonic for all N except $N = 4$, for which it is only monotonic on $p \in [0.1, 1.0]$). Perhaps surprisingly, however, the senders exhibit a more complex trend. After a slight increase in the probability of manipulating f_1 as p increases, the probability declines to a quantity far below the baseline. Both the initial increase and the final decrease is once again larger for greater N. We see, then, that despite an initial attraction of senders to f_1 with low-but-nonzero values of p, miscoordination becomes more common as p approaches 1.

The presence of possible miscoordination suggests further analysis to get a clear picture of how the conditional correlation of a channel influences the success of those senders and receivers who actually do coordinate on it. Figure 7 plots two natural statistics for the groups (in each condition) of agents who successfully coordinate on f_1. The first (left) is the alignment with nature. The effect we observe is that, as p increases, alignment goes up, but then it falls back (although still above the baseline at $p = 0.0$) when p reaches 1. This effect is greater with larger N, but peaks earlier: at $N = 2$ the peak value is at $p = 0.6$, at $N = 4$ it is at 0.7, at $N = 8$ it is at $p = 0.5$, and at $N = 16$ it is at 0.4. The second statistic (right) is a ratio of accuracies calculated in the following way. Let G be the group that coordinates on f_1, and G^C its complement. Then,

$$\text{Accuracy Ratio} = \frac{\text{Mean Cumulative Accuracy } (G)}{\text{Mean Cumulative Accuracy } (G^C)}. \tag{4}$$

The accuracy ratio reveals that, although the accuracy of players in aggregate is declining for middling values of p, it is actually increasing for those players who successfully coordinate on f_1. And, although the accuracy of players in aggregate increases toward 1 for high values of p, the accuracy ratio decreases to below 1. Further analysis reveals that this is not because the accuracy of the players who successfully coordinate on f_1 decreases, but because of the increase in accuracy of those who do not. When we examine the accuracy ratio when G is the group who coordinates on f_0 (the conditionally uncorrelated channel), a nearly-identical picture emerges, except that neither the proportional increase nor decrease is as extreme.[18] So, it is largely the coordination simpliciter, and not the coordination

[18] More specifically, for increasing N, the accuracy ratio peaks at $1.03, 1.05, 1.10, 1.19$ respectively, and takes on minimum value (at $p = 1.0$) of $1.00, 0.98, 0.95, 0.92$ respectively.

on the conditionally correlated channel in particular, that is having the largest effect on accuracy.

Appendix C. Model A/2 Results

Figure 8 compares the mean accuracy of agents in Model A/2 with the mean accuracy of agents in Model A/1 with $N = 2$. Note that the Model A/2 accuracy is everywhere higher, but exhibits the same trend. Namely, the accuracy is slightly lower at middling values of p than at the extremes. The difference in mean accuracy between the middle values and the extremes is very small: 0.996 at $p = 0.0$, to a low of 0.978 at $p = 0.4$, to a high of 0.998 at $p = 1.0$. But without reading tea leaves, no clear trend is visible from $p = 0.2$ to 0.9. Rather than representing an overall decrease in accuracy, the mean is being dragged down by a few agents who fail to coordinate successfully. At $p = 0.4$, for example, 0.003 of the sender/receiver pairs did not exceed an accuracy of 0.75. The lowest-scoring of these pairs had a cumulative accuracy of only 0.510.

Figure 9 compares the mean sender and receiver attentions for Models A/1 with $N = 2$ and A/2. The sender manipulation quantity this time is specifically how often the sender will manipulate f_1 when the predator is present. The first thing to notice is that when $p = 0.0$, the probability that a sender learns to manipulate f_1 to signal the presence of a predator is only 0.401, while the receiver learns to attend to f_1 0.502 of the time. The sender manipulation quantity is not a spurious statistical deviation from 0.5. To see how this works, consider all the possible signalling systems that could develop:

(1) manipulate f_0 when predator present, f_1 when absent
(2) manipulate f_0 when predator present, do nothing when absent
(3) manipulate f_1 when predator present, f_1 when absent
(4) manipulate f_1 when predator present, do nothing when absent
(5) do nothing when predator present, manipulate f_0 when absent
(6) do nothing when predator present, manipulate f_1 when absent

For (1), (2), and (5), if the receiver attends to f_0 perfect communication is possible. For (3), (4), and (6), perfect communication is possible if the receiver attends to f_1. There are thus 6 separating equilibria in the game. But, in only 2 of them does the sender manipulate f_1 when the predator is present. This partially explains why the senders learn to manipulate f_1 significantly less than half of the time, although more work is needed to determine why the learning dynamics select those equilibria more than one-third of the time.

Note, however, the broadly similar pattern present in sender manipulation and receiver attention when compared with the results in Model A/1. Senders initially become more likely to manipulate f_1 as p increases, before dropping back down as p approaches 1.0. Receivers gradually become more likely to attend to f_1 as p increases. The differences are that senders become most likely (0.611) to manipulate f_1 at $p = 0.7$, which is a peak at a higher value of p than for any case in Model A/1. Likewise, the receivers do not become significantly more likely to attend to f_1 until after $p = 0.7$, where the probability that a receiver will learn to attend to f_1 is 0.522. Nonetheless, the same basic patterns hold which demonstrate a divergence between the pragmatic success of agents and their success in communicating.

Appendix D. Model B Results

Figure 10 shows results for Model B. As with Model A, when there is no natural salience ($r = 1$, thus making it a vanilla $16 \times 16 \times 16$ signalling game), the mean alignment with nature is $0.062 \approx 1/16$, and the mean accuracy is 0.908. Here, the alignment with nature is based on how much the sender's evolved signalling dispositions match the signalling system which would pay off r (if the receiver's dispositions matched):

$$\text{(5)} \quad \text{Alignment} = \frac{\#\text{ of sender urns with modes corresponding to act paying off } r}{N}.$$

These quantities jump when $r = 10$ to a mean alignment of 0.476 and mean accuracy of 0.953, but subsequently level off. When $r = 1000$ the mean alignment is still only 0.481 and mean accuracy is only 0.959.

Appendix E. Model C Results

The formula which captures the average information transmitted by a sender in a signal when there are N equiprobable states of nature is:

$$\text{(6)} \quad \text{Avg Inf} = \sum_{\text{signals}} \sum_{\text{states}} P(\text{signal}) P(\text{state} \mid \text{signal}) \log[N \cdot P(\text{state} \mid \text{signal})]$$

Figure 11 plots the average information transmitted by senders in each of the conditions. A note on the figure: each point represents the average information transmitted by the sender at the end of a completed simulation. These are rank-ordered, so that they form an empirical CDF: the y-value of a point indicates the proportion of simulations for that condition in which senders transmitted at most the quantity of information specified on the x-axis. We can calculate the breakpoints for different ways in which a 2-term signalling convention can divide a 10-state world. If each term represents 5 states of nature (represent this as $\langle 5, 5 \rangle$), the average information transferred will be 1.0 bits. If one term represents 4 states and the other 6 states ($\langle 4, 6 \rangle$), average information will be 0.971. If $\langle 3, 7 \rangle$, 0.881. If $\langle 2, 8 \rangle$, 0.722. And finally, if $\langle 1, 10 \rangle$, 0.469.

When $r = 1$, the experiments replicate LaCroix's results. All agents end up with cumulative success rates greater than 0.9. We can approximate how many agents ended up with each kind of language by examining how many senders' average information transfers lie between two breakpoints. If a sender has average information between two breakpoints, then we will assume for simplicity that the agent's language is heading towards the partition with higher average information. This replicates LaCroix's results, giving:

232 agents with $\langle 5, 5 \rangle$ partitions.
524 agents with $\langle 4, 6 \rangle$ partitions.
205 agents with $\langle 3, 7 \rangle$ partitions.
38 agents with $\langle 2, 8 \rangle$ partitions.
1 agent with $\langle 1, 9 \rangle$ partitions.

When $r = 2$, we already see significant deviation from the behaviour when $r = 1$. Once again, all agents have cumulative accuracy greater than 0.9. But, fewer agents end up with languages that partition the world equally. We have:

64 agents with $\langle 5, 5\rangle$ partitions.
441 agents with $\langle 4, 6\rangle$ partitions.
381 agents with $\langle 3, 7\rangle$ partitions.
100 agents with $\langle 2, 8\rangle$ partitions.
14 agents with $\langle 1, 9\rangle$ partitions.

Finally, when $r = 10$ we see many agents learning the $\langle 1, 9\rangle$ and $\langle 2, 8\rangle$ partitions. 997 agents end up with cumulative accuracy greater than 0.19, but 3 do not. These have accuracy very close to 0.10: they did not learn a term language at all. Of the other 997:

1 agent with $\langle 5, 5\rangle$ partitions.
11 agents with $\langle 4, 6\rangle$ partitions.
146 agents with $\langle 3, 7\rangle$ partitions.
476 agents with $\langle 2, 8\rangle$ partitions.
363 agents with $\langle 1, 9\rangle$ partitions.

Acknowledgements

The authors thank Jeffrey Barrett, Simon Huttegger, Nathaniel Imel, Brian Skyrms, and the participants at the Game Theory paper session at PSA 2022 for helpful feedback. Thanks also to Mark and Martha VanDrunen for their hospitality while the draft was being written.

Jacob VanDrunen
Department of Logic and Philosophy of Science
University of California, Irvine
Irvine, CA, USA
jvandrun@uci.edu

Daniel A. Herrmann
Department of Logic and Philosophy of Science
University of California, Irvine
Irvine, CA, USA
daherrma@uci.edu

References

Alexander, J. M., Skyrms, B., and Zabell, S. L. (2012). Inventing new signals. *Dynamic Games and Applications*, 2(1):129–145. Publisher: Springer.

Argiento, R., Pemantle, R., Skyrms, B., and Volkov, S. (2009). Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic processes and their applications*, 119(2):373–390. Publisher: Elsevier.

Barrett, J. A. (2006). Numerical Simulations of the Lewis Signaling Game: Learning Strategies, Pooling Equilibria, and the Evolution of Grammar. Technical Report MBS 06-09.

Barrett, J. A. (2021). Self-Assembling Games and the Evolution of Salience. *The British Journal for the Philosophy of Science*. Publisher: The University of Chicago Press.

Barrett, J. A. and Skyrms, B. (2017). Self-assembling Games. *The British Journal for the Philosophy of Science*, 68(2):329–353. Publisher: The University of Chicago Press.

Barrett, J. A. and VanDrunen, J. (2022). Language Games and the Emergence of Discourse. *Synthese*.

Cubitt, R. P. and Sugden, R. (2003). Common knowledge, salience and convention: A reconstruction of David Lewis' game theory. *Economics & Philosophy*, 19(2):175–210. Publisher: Cambridge University Press.

Darwin, C. (1875). *The Descent of Man, and Selection in Relation to Sex*. D. Appleton and Company, 2nd edition.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66(3):377–388. Publisher: [Duke University Press, Philosophical Review].

Herrmann, D. A. and VanDrunen, J. (2022). Sifting the Signal from the Noise. *The British Journal for the Philosophy of Science*. Publisher: The University of Chicago Press.

Hu, Y., Skyrms, B., and Tarrès, P. (2011). Reinforcement learning in signaling game. *arXiv:1103.5818 [math]*. arXiv: 1103.5818.

Huxley, J. S. (1966). A discussion on ritualization of behaviour in animals and man. *Philosophical Transactions of the Royal Society of London B*, 251:249–271.

Kavanagh, M. (1980). Invasion of the Forest By an African Savannah Monkey: Behavioural Adaptations. *Behaviour*, 73(3-4):238–260. Publisher: Brill.

Kelley, L. A., Coe, R. L., Madden, J. R., and Healy, S. D. (2008). Vocal mimicry in songbirds. *Animal Behaviour*, 76(3):521–528.

LaCroix, T. (2020a). Communicative bottlenecks lead to maximal information transfer. *Journal of Experimental & Theoretical Artificial Intelligence*, 32(6):997–1014. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0952813X.2020.1716857.

LaCroix, T. (2020b). On salience and signaling in sender–receiver games: partial pooling, learning, and focal points. *Synthese*, 197(4):1725–1747.

Lewis, D. K. (1969). *Convention*. Harvard University Press, Cambridge, 2002 reprint edition.

Lewis, J. (2009). As well as words: Congo Pygmy hunting, mimicry, and play. In *The cradle of language*, page 236. Oxford University Press.

Lorenz, K. Z. (1966). Evolution of ritualization in the biological and cultural spheres. *Philosophical Transactions of the Royal Society of London B*, 251(772):273–284. Publisher: Royal Society.

Planer, R. and Sterelny, K. (2021). *From signal to symbol: the evolution of language*. MIT Press.

Quine, W. V. (1936). Truth by Convention. In *Philosophical Essays for Alfred North Whitehead*, pages 90–124. London: Longmans, Green & Co.

Ratnayake, C. P., Goodale, E., and Kotagama, S. W. (2009). Two sympatric species of passerine birds imitate the same raptor calls in alarm contexts. *Naturwissenschaften*, 97(1):103.

Rousseau, J.-J. (2011). *Rousseau: The basic political writings*. Hackett Publishing, 2nd edition.

Scott, J. L., Kawahara, A. Y., Skevington, J. H., Yen, S.-H., Sami, A., Smith, M. L., and Yack, J. E. (2010). The evolutionary origins of ritualized acoustic signals in caterpillars. *Nature Communications*, 1(1):4. Number: 1 Publisher: Nature Publishing Group.

Skyrms, B. (2000). Evolution of inference. In *Dynamics of human and primate societies*, pages 77–88. Oxford University Press.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. Transaction Publishers. Google-Books-ID: Go8XozILUJYC.

Tinbergen, N. (1952). "Derived" Activities; Their Causation, Biological Significance, Origin, and Emancipation During Evolution. *The Quarterly Review of Biology*, 27(1):1–32. Publisher: The University of Chicago Press.

Tucker, M., Shah, J., Levy, R., and Zaslavsky, N. (2022). Towards Human-Agent Communication via the Information Bottleneck Principle. Technical Report arXiv:2207.00088, arXiv. arXiv:2207.00088 [cs] type: article.

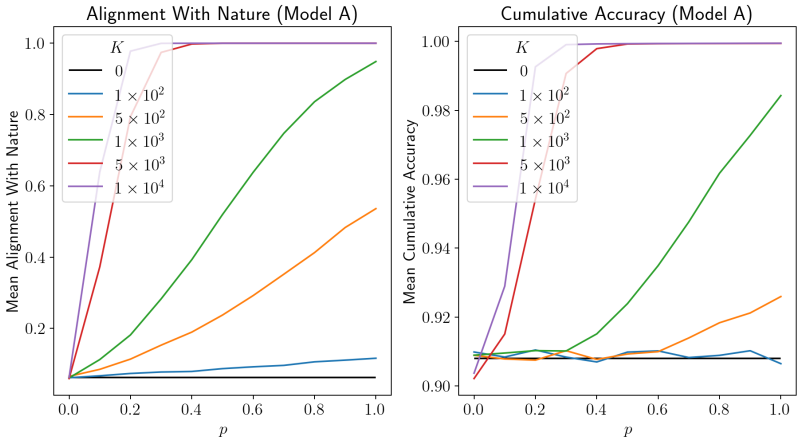

FIGURE 4. Results for Model A showing the alignment of the final signalling convention with the mapping given by the initial correlation of the signalling channel (left), and the cumulative accuracy of the players after 10^7 plays (right). Means are taken out of a sample size of 10^3 simulations for each condition.

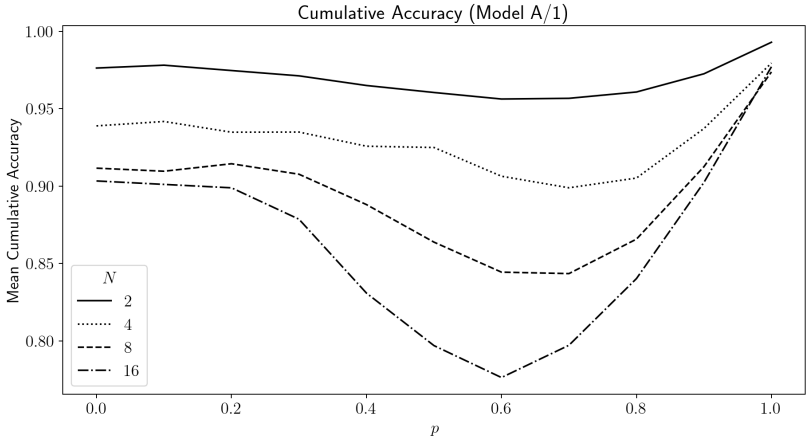

FIGURE 5. Results for Model A/1 showing the cumulative accuracy of the players.

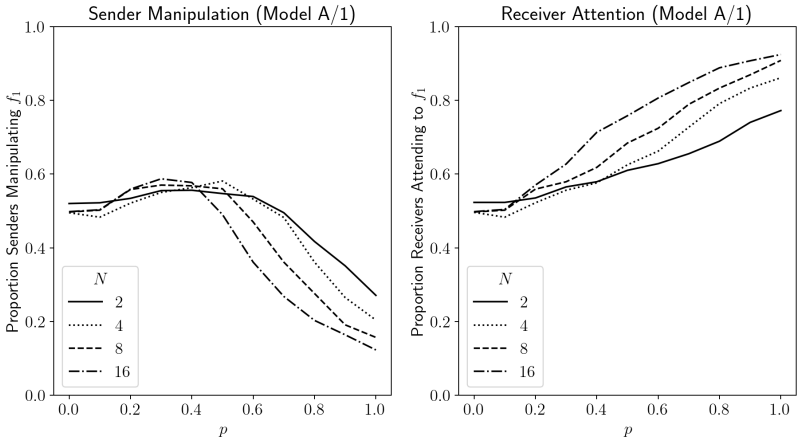

FIGURE 6. Results for Model A/1 showing the proportion of senders (left) and receivers (right) attuned to the conditionally correlated channel $\boldsymbol{f}_1$.

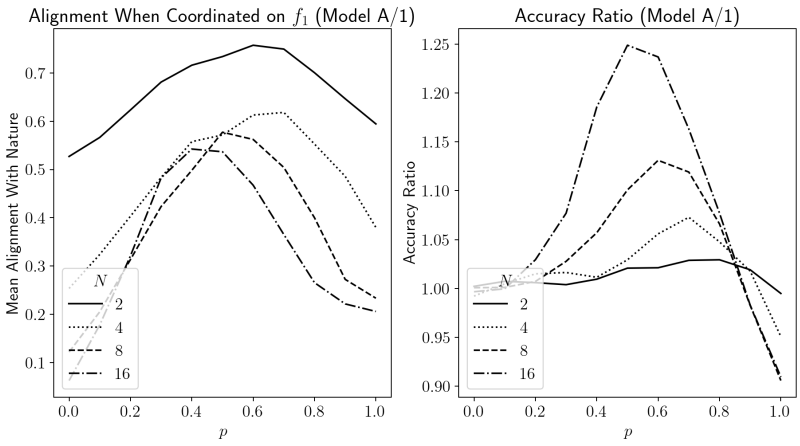

FIGURE 7. Results for Model A/1 showing the alignment of the final signalling convention when both sender and receiver are coordinated on $\boldsymbol{f}_1$ (left), and the ratio of the accuracy of players when coordinated on $\boldsymbol{f}_1$ to the accuracy of players under all other outcomes.

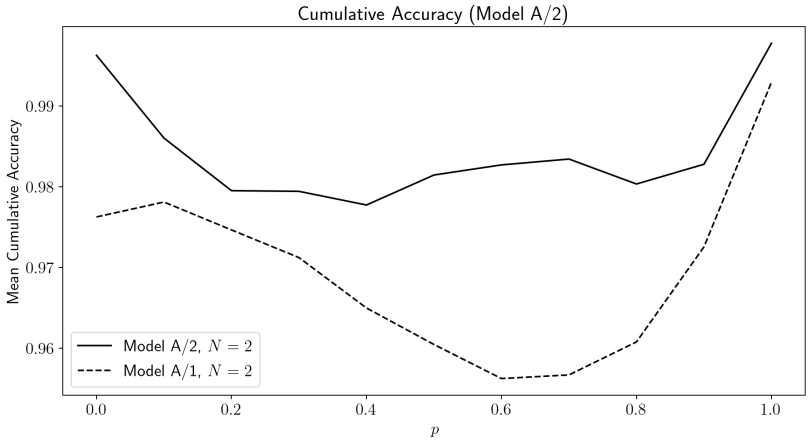

FIGURE 8. Results for Model A/2 showing the cumulative accuracy of the players, compared with the accuracies of players in Model A/1 with $N = 2$.

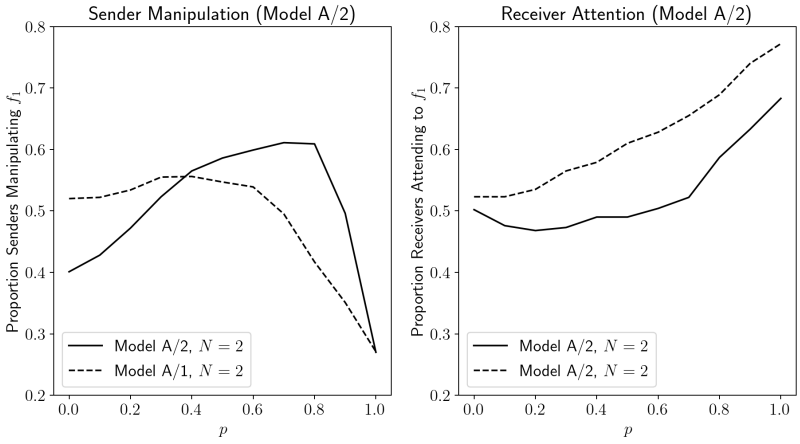

FIGURE 9. Results for Model A/2 showing the proportion of senders (left) and receivers (right) attuned to the conditionally correlated channel f_1, compared with the players in Model A/1 with $N = 2$.