December, 2018

Conventionality and Causality in Lewis-Type Evolutionary Prediction Games

Gordon Purves, *Sacred Heart University*

Conventionality and Causality in Lewis-Type Evolutionary Prediction Games

Gordon Michael Purves

Abstract

Barrett and others have used Lewis-style evolutionary games to argue that we ought not to trust our scientific languages to inform us about ontology. More specifically, Barrett has shown that in some simple evolutionary contexts the best descriptive languages need not cut nature at its joints, that they may guide action as successfully as possible while simultaneously being deeply conventional. The present article expands upon Barrett's argument, exploring the space for conventionalism in more metaphysically robust causal evolutionary models. By using evolutionary prediction games, the present article examines the range of metaphysical facts in these models and shows how Barrett's sceptical argument may be both challenged and expanded to cover these systems, resulting in a more thoroughgoing scepticism about what metaphysical conclusions may be drawn from linguistic and theoretical structure.

1 Introduction

One of the more entrenched debates about the proper method of metaphysical inquiry concerns the role that our ordinary and scientific languages are to play in the construction of metaphysical hypotheses. The question, here, is in part whether we should trust that the kinds picked out by the terms of our languages (whether natural or scientific) track real natural kinds, or whether they are simply pragmatically useful fictional constructions. As a simple example, we may ask whether the fact that we pick out a group of objects under the broad category of 'trees' or 'mammals' should lead us to believe that these categories approximate real ones that exists in a metaphysically important way antecedently to our naming them, or whether such a category is simply a useful way for us to group some apparently similar objects together. Put more broadly, we may ask what we may deduce about the fine structure of the world from the success of our empirical descriptions of it.

Along these lines, Goodman ([1983]) famously argued that how entrenched a particular term language is may allow one to distinguish between natural and artificial kinds. He concludes that we are justified in preferring the categories 'blue' and 'green' over the categories 'grue' and 'bleen' because of the former's history of use in successful predictions. On the other hand, Barrett ([2007]) proposed that we can analyse the trustworthiness of entrenchment for metaphysics by constructing evolutionary signalling games of the sort that Lewis ([1969, 2002]) devised. The idea, here, is that we can mathematically model a simple evolutionary system as a game within which the players may learn to cooperatively construct a language to describe the 'world' around them. Barrett shows that, under certain conditions, the linguistic kinds that the players create are conventional, even when the signalling system that evolves is as effective at guiding action as possible. Barrett takes these results to suggest that the conventionality of language may run deep, and that, therefore, we ought not to take the fact that natural and scientific languages are very good at guiding our interactions with the world to be a reason to think that they are cutting nature at its joints. Barrett's argument is a sceptical one. He is not arguing that the kinds in our natural and scientific languages are conventional or fictional so much as he is arguing that their success is not evidence that are not merely conventional.

Barrett's argument is partially an artefact of the specific game that he is using insofar as his model does not have any metaphysical facts or kinds that could be captured by the evolved signalling system. This is not a problem for the sceptical argument, since he is showing that a signalling system may have a 'kind language' without there being any natural kinds at all, not that there could not be such natural kinds. Nonetheless, it leaves open the question of whether or how the existence of natural kinds or metaphysical structure more generally might affect the linguistic structures that can evolve. Put another way, there remains the possibility that, though it is surely possible that there are no natural kinds at all and that our languages are conventional all the way down, we can nonetheless make assertions along the lines of, 'if there are real metaphysical facts, then the success of our languages tells us something about how some of them must be'. The present article seeks to address this possibility more directly by exploring the possible metaphysical interpretations available in more complex Lewis-type signalling games that have a deeper structure available to be discovered. Ultimately, it will be argued that even these more modest, hypothetical conclusions are unsupported by the empirical success of the evolved language. This will be accomplished by showing that, even when there are real metaphysical facts available to be discovered, nonetheless the propositions of such a hypothetical metaphysics cannot be separated from their fictional cousins.

I will begin by briefly introducing the reader to Lewis signalling games, and will rehearse the most salient points of Barrett's argument, interpreting it in light of the considerations noted above. Following that, I will introduce the core idea of a prediction game, and will show that, unlike in the 'worlds' of Barrett's signalling game, there are substantive metaphysical facts that are (more or less) expressible in an evolved signalling system and are emergent from an analysis of the structure of that system. Finally, by comparing multiple evolved signalling systems, I will show that these true metaphysical facts cannot be separated from fictional propositions that emerge along with them. As a result, scepticism along the lines that Barrett proposes may be applied to more speculative and hypothetical metaphysical inquiry as well.

2 Lewis-Style Games

A Lewis-type signalling game (Lewis [1969, 2002]), as re-imagined by Skyrms ([2008]), models the collective acquisition of knowledge through the stochastic development of rudimentary language. A simple n-state/n-term game models a world that has, at any given moment, one of n possible states, and that periodically evolves to a random new state. The game has two players: the sender and the receiver. On each play of the game, the sender observes the state of the world and selects one of n possible signals. The receiver hears the signal (without observing the world) and chooses one of n possible actions to perform based on it, where each action corresponds to a state of the world. If the action chosen matches the state of the world, then both sender and receiver are 'rewarded', meaning that the probability distribution that determines which of the signals and which of the acts are performed in each case are modified such that the successful combination becomes slightly more likely.

There are many ways that this basic structure may be modified or implemented, some of which will be explored below, but the present work will restrict itself to a simple reinforcement learning algorithm. With this algorithm, the sender and receiver can be imagined to have bins corresponding to each possible input, each filled with a number of balls that represent the different outputs. An output is picked by randomly selecting a ball from the bin of the corresponding input, and if the sender's input matches the receiver's output, then another ball of the type each used is added to their respective input bins. For example, if the world evolves to state two, then the sender will choose a signal ball at random from bin two. If the sender chooses signal A, then the receiver will go to bin A and choose an act ball at random. If the receiver chooses act two from bin A, then both players are rewarded by adding an additional A ball to the sender's bin two and a two ball to the receiver's bin A. This process is repeated for many runs, with the world selecting a state (sender input), which probabilistically determines a signal (sender output / receiver input), which then probabilistically determines an act (receiver output), which will be reinforced whenever it correctly matches the original state of the world. In addition to having the virtue of being easy to model, this algorithm, as Barrett ([2007], p. 531) points out, has a long pedigree in the psychological literature and tracks closely the replicator dynamics of natural selection (Huttegger [2007a], [2007b]; Skyrms [2009]). In other words, though this model is, obviously, extremely abstract and simple, it nonetheless is a reliable model of real evolutionary systems. Furthermore, while there are other models that have proven themselves to be more robust, for instance (Alexander [2014]), the more complex algorithms do not diverge from simple reinforcement learning in any ways that are relevant to the present argument.

Under some fairly reasonable constraints, several researchers (Skryms [2008], [2010]; Barrett [2007]; Argiento *et al*. [2009]; Huttegger *et al*. 2010]) have shown that the sender and receiver will usually develop a reliable signalling system without any initial salience needed between signal and state. This is in direct contradiction with the classic intuitive accounts of the origin of language, according to which some sort of natural salience (for instance onomatopoeia, or an intuitive understanding of natural kinds) is a necessary precondition for the evolutionary development of language. (For a recent defence of a position similar to this, see (Larsson [2015]).) Simple Lewis-style signalling games may, therefore, be taken to prove that purely conventional languages can arise from a process of natural selection. Moreover, in slightly modified signalling games the signalling system that emerges may be syntactically, as well as semantically, complex (Barrett [unpublished], [2007], [2014b]).

3 Barrett's Conventionality

Barrett ([unpublished], [2007], [2014b]) argues that the conventionality of meaning exhibited by Lewis-style games can be extended to include a conventionality about how the evolved signalling system cuts up the 'natural' world. It is easy to construct a variant of the original Lewis signalling game to illustrate this. Suppose, for example, that the world has four possible states, and four corresponding acts, but our language only has two terms. It is then impossible for perfect signalling to arise, since there are not enough terms to encode all of the information in the world. What happens instead is the sender associates several states with a single term (effectively grouping them as a 'kind'), and the actor will select an action that will match a random member of the group of states picked out by that signal.

This could happen in many different ways. If we name the states {1, 2, 3, 4}, and the terms {A, B}, then the sender may use A to indicate that either state one or two has occurred, and B to indicate either state three or four. Whenever the receiver hears signal A, then, they will pick at random between act one and act two, and similarly when they hear signal B. The players can, in this scenario, expect to succeed ½ of the time, a significant improvement over the ¼ success rate they would have if actions were chosen completely at random. But this is far from the only way things could end up. They could just as easily use A to pick out {1, 3}, and B to pick out {2, 4}. Or, A could pick out {1}, and B could pick out {2, 3, 4}. Or we could have any permutation of these results. Each of these pooling equilibria have a success rate of ½, and thus mark a real improvement upon the original random signalling, that has a success rate of ¼. Moreover, since they are equally successful adaptations, none is an improvement upon any other and they are equally likely results. As such, the kinds picked out by each language have little connection to the kinds picked out by the world, though the simple term languages are as successful as they could possibly be.

This partly stems from the equiprobability of the states. Suppose, for instance, that states one and three each occur 40% of the time and two and four occur 10% of the time.

It is now no longer true that each equilibrium is equally adaptive. An equilibrium that groups one with three and two with four signals successfully 50% of the time. On the other hand, if one is proportionally grouped with two (that is, the receiver performs act one 80% of the time it receives that signal, and act two otherwise), and the same for three with four, then the receiver act correctly 68% of the time. This isn't much of a problem for two reasons. First, since the 1–2, 3–4 grouping is as successful as the 1–4, 2–3 grouping, and neither captures true metaphysical categories, much conventionality remains. Further, all that is necessary for this emergent conventionality is that some joint sets of states are roughly equiprobable, a condition that holds more often as the number of states increases. As such, while there is an artificiality to the equiprobability assumption, it models well the fact that more realistically complex models usually have some approximately equiprobable sets of states. Second, the aim of the game is sceptical, and so it need only demonstrate the possibility of deep conventionality. As will be discussed in greater detail below, since the players of the game do not have access to other possible evolutionary outcomes or other state distributions, they are unable to determine whether their signalling system's structure is true metaphysics or convention.

Of course, it remains true that part of why we have a variety of possible ontological conventions is the fact that the sender is not able to completely describe the world. This need not concern us here for a few reasons. First, it is entirely reasonable to suppose that, in nature, there are not enough terms available to any organism to perfectly describe any state of the world; indeed, given the complexity of nature, it would be a miracle were any creature to evolve to be able to completely describe the world around it in a simple term language. O'Connor ([forthcoming]) has explored this act of conceptualization explicitly in the context of Lewis-style games, and concludes with a conventionalist scepticism that is highly compatible with that defended here.

Moreover, there is already a deep and realistic salience built into any such game, insofar as the world is simplified to just a few states. If this were to describe the full extent of the world, then, pretty obviously, it would be too simple for creatures of any kind to evolve. What this really means is that there are only a few states that matter to the players of the game, and, as O'Connor ([forthcoming]) has shown, there is conventionalism about the creation of these states-of-interest even after interests are fixed. Thus, even if they have the linguistic resources to describe each state within the game, that, in fact, says more about the players than about the world. Consider a real-world example that has become common within the philosophical literature: vervet monkey signalling. Vervet monkeys have been observed to have dozens of different alarm calls with a variety of different levels of specificity, but three are particularly well documented (Seyfarth *et al.* [1980]). One warns of an approaching leopard, and causes the monkeys to run up a tree to the small branches that a leopard cannot reach. Another warns of a nearby eagle, and elicits a response of looking up and running for the nearest bush. A third alerts them to the presence of a python, and causes them to stand tall on two feet and look at the ground. This signalling system is highly successful, in part, because it is overly simple and ignores much information that they could attempt to transmit. The signalling system does not suggest that there are only three possible states of the world (or, in the less simplified and abstracted case, a few dozen); rather, it says that these are the most important states for the continued survival of the species and of individuals within the species. Thus, a vervet metaphysician (if I may be allowed a little latitude) could conclude very little about the vervet-independent world from the success of their signalling system. Rather, insofar as it is successful, the signalling system only indicates that there are a few things that are particularly important for vervet monkeys and they have become adept at communicating the presence of exactly those things very efficiently. More generally, Barrett argues that what constitutes success for a signalling system says more about the signallers than it does about the fine structure of the world, and thus is reflective of the deep conventionality inherent to languages of any sort.

Finally, and most importantly for the present discussion, another conventionality of kinds can arise in some cases where there are enough linguistic tools available to be completely successful. The key to this leap is to switch from simple semantic games, within that a single signal is sent with each play of the game, and thus ideally there is a one-to-one correspondence between the states of the world and the terms available to describe it, to what Barrett terms syntactic games, within that the language is free to evolve a simple grammar.

Suppose that our sender is able to send two signals for each run of the game, such that the signals are independent of each other. We could think of this as having two signallers who cannot communicate to each other, or as a single signaller sending two completely independent signals, one after the other. What is most important is that the two signals are separate and independently chosen, and thus cannot be thought of as ‘syllables’ of the same term. The receiver can hear and differentiate between both signals, and then will proceed to pick a single act that may (or may not) match the state of the world.

Keeping with the four-state, four-act world discussed above, and only allowing two terms for each of the two signals, A and B for the first term and X and Y for the second, we can easily parse the world completely by grouping multiple states for each term. For instance, the players could evolve to categorize states one and two together as A, states three and four as B, states one and three as X and states two and four as Y. This would generate a complete signalling system, since any pair of terms would uniquely pick out a single state (state one would correspond to the signals AX, state two would correspond to signals AY, and so on).

What is important here is, first, that the two signals are independent of each other and, second, that every permutation of these results will be an equally successful signalling system.[1] The independence keeps the two signals distinct, making them unique terms, and as a result the variety of equally successful signalling systems is a variety of equally successful ways of parsing the world into linguistic kinds that arbitrarily group states of nature. In other words, Barrett concludes, if our language could be equally successful by grouping states one and two together as a single term as it could grouping any other pair, then the conventionality that picks out the name of a group extends to the ontology of the group itself. Thus, it is perfectly possible that an ideally successful

[1] Again, that every permutation is equally successful is because all states are equiprobable. If probabilities are different, there will be some permutations that are more adaptive than others, but this is only true if there are very few states. All that is relevant is that there are, here, some sets of states that may be equiprobable with some others, and that that results in multiple permutations being equally successful signalling systems.

descriptive language completely fails to cut nature at its joints, because it is entirely possible that there are no natural kinds to be identified. Barrett takes this conclusion to undercut the legitimacy of metaphysical inquiry that takes as its starting point how our natural or scientific languages divide the world.

Furthermore, though a complete signalling system is possible, it is not always achieved. Rather, the players will occasionally reach a point of 'partial pooling', a suboptimal state in which the learning algorithm gets 'stuck'.[2] This could happen if term A were to correspond to state one only (in which case, the second signal would be irrelevant), then term B would indicate that one of states two, three, or four has occurred. The second signal, then, would have to divide three states with only two terms. This could be achieved by having signal BX indicate state two and signal BY indicate that either state three or state four has occurred, in which case the receiver would choose its action at random between the two. Or, perhaps, signal BX indicates state two and signal BY indicates state three, but there is no signal that captures state four (and so a random signal is sent when state four occurs). Several studies have shown how to mitigate partial pooling, but it can never be fully eliminated except in the two-state, two-term, two-act game (Argiento *et al*. [2009]; Huttegger *et al*. [2010]). This further undercuts the reliability of our linguistic systems' ability to provide a 'true' ontology, and seems to leave us with a highly fallible nominalism.

Barrett ([2007], pp. 543–4) concludes:

[2] This is not unique to syntactic games, and is reasonably common in any game more complex than two-state, two-term, two-act signalling games; see (Huttegger *et al*. [2010]).

> While [the players] may try to infer the structure of the world from their best descriptive language, their philosophical reflections may not yield at all what they want. What they can reliably infer from their successful kind language is that it is in reflective equilibrium with the dispositions that evolved it, and what they may have been tempted to take as natural kinds are purely conventional partitions of the state space contingent on the available linguistic resources and the dispositional reinforcements of the game in which the language evolved.

Put in other words, though there is a clear grouping of objects into kinds, and that grouping may lead to successful action guidance, nonetheless the grouping is completely conventional, and says nothing more about the world than does our decision to use the sign 'tree' to refer to a particular group of objects.

This argument does not, however, say much about the space of metaphysical theories that are possible, other than that there is a theory that cannot be ruled out, namely the theory that there is no metaphysical structure at all. It may still be the case that some information can be gained about other metaphysical claims, other interpretations of the success. Such a view is akin to the science student who, upon learning of Cartesian scepticism, responds that, while it is possible that an evil demon has fabricated all of our experiences, and true that the success of science cannot eliminate or reduce that chance, nonetheless good scientific inquiry can say a great deal about how the world is, or would be, supposing that we are not actually being deceived by an evil demon, are not brains in a vat, and are not being fooled by any similar tool of radical scepticism.

Thus, though Barrett's argument achieves its sceptical goal, it is also limited. These syntactic games do not suffice to justify the conclusion that we ought to think of our natural and scientific language systems as deeply conventional. Rather, they show that the ability of these languages to successfully guide action is not a sufficient reason to conclude that they are not deeply conventional. Of course, the signalling systems in the games that Barrett considers actually are deeply conventional, but that is, in part, because there are no natural kinds to latch on to. It remains conceptually possible that, while in worlds analogous to Barrett's games a conventionality of kinds does arise, and we cannot prove that our world isn't one of them, nonetheless when there are real metaphysical facts to latch onto, the evolved language may capture some of those facts in a way that is not deeply conventional. Barrett has shown that we cannot know that we have latched on to those facts, but has stopped short of showing that we cannot know what those facts would be if they existed. Furthermore, such claims are far from abnormal; naturalistically inclined philosophers often begin a metaphysical project by interpreting scientific theories, for instance by saying things like, if the general theory of relativity is approximately correct, then space and time may be like this or like that, but not some other. As a result, Barrett's sceptical conclusion is not quite as sweeping as it first appears.

For an inference from linguistic facts to metaphysical facts to be sound, two conditions must be met. First, there must be some true, non-conventional facts of a metaphysical nature that are more or less expressible within the language. Second, these facts must be identifiable and separable from the purely conventional elements of the conceptual scheme. In the next section, I will introduce the prediction games described by (Barrett *et al*. [unpublished]; Barrett [2014a]; Dickson [2013]), which, by virtue of having a causal structure, include the possibility of true metaphysical claims. In section four, I will demonstrate that the first condition is met in this game by showing that versions of those metaphysical claims are emergent from the structure of an evolved signalling system in a way that holds true across both ideal and non-ideal evolved signalling systems. In the final section, however, I will argue that the second condition is not met, because the true emergent facts about the world are indistinguishable from within a linguistic system from the conventional emergent facts about the signalling system itself. I will conclude, then, that scepticism about any inference from linguistic structure to metaphysics, regardless of whether the inference is conditionalized or not, is warranted.

4 Evolutionary Prediction Games

In a prediction game, rather than being purely random, the evolution from one state to another is governed by probabilistic causal laws, and the goal of each play of the game is not to describe the current state, but to predict the next state. A successful signalling system does not simply parse the world into terms that they reliably identify, but must also successfully track the causal laws that govern the world. This modification may seem trivial, and if the causal laws are deterministic and non-stochastic, then the resulting game is generally uninteresting, as Barrett notes (Barrett [2014b]). The interesting cases come when the laws are non-deterministic or stochastic.

Consider a simple example of a four-state, four-term, four-act prediction game with a basic causal ring: According to this causal system, state one is usually followed by state two, state two by three, state three by four, and state four by one. The likelihood of the states evolving according to this trajectory at each turn is 90%, and if it does not follow the law, then there is an even probability of any state being next (including the next state, so there is actually a 92.5% probability of the next state in the sequence). If we define a complete run of the game as 500,000 plays, then partial pooling occurs roughly 16.9% of the time in this game. This result is a slightly (though consistently) higher rate than the analogous four-state, four-term, four-act signalling game, which has partial pooling roughly 15.9% of the time.

(It is, perhaps, worth noting that I am measuring partial pooling differently than Barrett and most other authors in the field, who generally use a direct measure of information content, the proportion of successes to total plays, or the time to convergence. Rather than these measures, I am checking, at the end of a complete run, if any of the bins have more than 15% of the balls of a type other than the largest. The 15% rate ends up being unproblematically arbitrary. If pooling occurs, it usually does so by splitting between two or, very rarely, more ball types close to evenly; if pooling doesn't occur, rarely more than 2% of the balls are of a minority type. There are, of course, much more exact measures available, but the added complexity and computing time is unnecessary for the present argument, though it does very slightly affect the measured success rates as compared to the other works cited above. For instance, (Barrett [unpublished]) reports 21.9% failure rate for the $4 \times 4 \times 4$ model, significantly higher than the 15.9% I am reporting. The difference arises because total average success may be below his threshold, even if the partial pooling is, after a few hundred thousand plays, resolved into a signalling system that, from that point on, has a much higher success rate. These differences largely disappear, furthermore, when the adjustments considered below—persistent local experimentation and gradual forgetting—are implemented, since the past failures to converge largely disappear with the addition of those resources. In any case, the reason to focus on directly measuring the partial pooling rate, rather than the success rate, is both to avoid these false negatives and so that the indetermination of the evolution of the prediction game does not count as failure, and, therefore, signalling games may be more directly compared to prediction games.)

As reported in (Barrett *et al.* [unpublished]), the partial pooling rate can be decreased by implementing what (Roth and Erev [1995]) call 'persistent local experimentation' and 'gradual forgetting' (see also Barrett and Zollman [2009]). The present implementation of these mechanisms departs slightly from that of these authors to ease computation times and to apply to the specific structure of this game, though the spirit and effect remain largely the same. 'Experimentation' here means that there is a small probability, ψ, at each decision (about what signal to send and about what act to perform), that the player will instead choose at random. Forgetting is implemented every N turns by multiplying the number of balls in each bin by a factor, ϕ, between zero and one. Both of these mechanisms can act to 'upset' the stability of partial pooling for a fairly wide range of parameters, and will be incorporated in the remaining models in the present work, though they serve more to make the results clearer than to fundamentally affect the argument itself.

5 Emergent Ontology

Now, obviously, this simple prediction game will demonstrate the conventionality of Lewis's original signalling game: the terms attach to the world in a purely conventional way. What is of new interest is a further expansion of the prediction game to incorporate syntax in the way that Barrett described. We will here use the same initial setup as Barrett: four states, four acts, but dividing the four terms between two different signals. As before, there is space in this system to construct a complete signalling system, and though there

Sender 1	Signal A	Signal B
State 1	99.5%	0.5%
State 2	99.4%	0.6%
State 3	0.6%	99.4%
State 4	0.6%	99.4%

Sender 2	Signal X	Signal Y
State 1	99.5%	0.5%
State 2	0.6%	99.4%
State 3	99.4%	0.6%
State 4	0.6%	99.4%

Receiver	Act 1	Act 2	Act 3	Act 4
Signal AX	0.5%	98.4%	0.5%	0.5%
Signal AY	0.5%	0.5%	98.4%	0.5%
Signal BX	0.6%	0.6%	0.6%	98.3%
Signal BY	98.4%	0.5%	0.5%	0.5%

Table 1. Probability of each signal, given a state, and each act, given a signal, for a run of 500,000 plays in a four-state, two-term, two-sender, four-act predictor game, with the causal law followed 90% of the time, and the forgetting factor $\phi = 0.90$, forgetting period $N = 400$, and the experimental factor $\psi = 0.01$.

is the possibility of partial pooling, a complete signalling system is usually achieved.

Consider a typical result (see Table 1):[3]

[3] Note that, though, for instance, Signal AX is sent when State one occurs, the receiver takes it to indicate that Act two is called for, since State two is likely to follow State one, and it is the next state that determines the appropriate act.

In this case, term A refers to the set of states {1, 2}, B to {3, 4}, X to {1, 3}, and Y to {2, 4}. Since there is no natural kind that corresponds to {1, 2} (or to any other grouping of states), the ontology of the emergent signalling system is conventional with regards to kinds just as with Barrett's model. The success of the kind language, again, indicates that it is in an equilibrium with the dispositions that evolved it, and thus the agents have successfully represented state-action pairs so as to reliably produce a positive payoff. Unlike in the syntactic signalling game, however, there are more facts in this world than whether a state-action pair produces a positive payoff. In the prediction game, there are additional facts about the causal structure and laws of the world.

Recall that the causal laws in this world may be summarized as follows: state one causes state two, state two causes state three, state three causes state four, and state four causes state two. None of these causes are absolute, since there is always a chance that a random state will follow, rather than the causally prescribed one; the world, thus is governed by a probabilistic causal ring. There are, therefore, in addition to the fact about the four natural kinds, five facts about the causal structure of the world: four laws and a more general ring topology. Of course, this understanding of the facts only makes sense from our position outside of the game. What really matters is what can be said from within the evolved signalling system.

In order to create a successful signalling system, the players have to have picked out some causal laws. There are a couple of ways that the causal laws, as grasped by the players, may be expressed. First, they may be understood as follows:

A and X (state 1)	→	A and Y (state 2)
A and Y (state 2)	→	B and X (state 3)
B and X (state 3)	→	B and Y (state 4)
B and Y (state 4)	→	A and X (state 1)

Alternatively, we may understand the causal laws through the individual terms:

- when X is true, neither A nor B will change truth values, but
- when Y is true, A and B both change truth values;
- when A is true, neither X nor Y will change truth values, but
- when B is true, X and Y both change truth values.

This latter expression of the laws pretty clearly differs a great deal from the actual causal laws (as understood, if I may be so bold, from the point of view of the world or from a god's eye view). Finally, the evolution of the world may be understood by looking at the pattern of the individual terms:

A → A → B → B, and X → Y → X → Y.

It seems clear that the causal laws as understood by the players—particularly in these latter two expressions—differs significantly from our original expression of how the laws were built into the world. Moreover, the plurality of ways of understanding these laws suggests another kind of convention itself. This convention, however, we may easily see is shallow in the sense that, though the differing interpretations clearly describe the world in different terms, a player within the game is capable of seeing and thus freely choosing between the three options. Put in other terms, a player of the game who wants to describe the causal structure of the world is forced to use one of these three interpretations, but they are positioned to be able to see that the three options are isomorphic, and thus the decision is largely irrelevant. This is akin to the clear conventionality involved in choosing to represent the world using the metric system or the imperial system, choosing to describe a scene in Spanish or German, or choosing between two equally adequate modelling methods. Of course, recognizing that these choices are shallow conventions does not resolve our initial question of whether it is possible, from within the linguistic scheme of the players, to identify the true causal structure of their world.

The shallowness of the conventional decision to use one of these three formulations of the causal laws gives us a cue for how to determine how deep the other conventions in this signalling system are. Specifically, any feature of the players' expression of the causal laws that, though different from the world's 'expression', is, nonetheless, the only way that the laws may be understood could provide an ontologically interesting insight into how the world is in itself, if that uniqueness is discoverable from within the game. Our task, then, will be to see what, if any, non-conventional metaphysical truths may be derived from the nature of the conventionality in the signalling system, and to determine whether the players of the game would be capable of learning these truths.

Some non-conventional facts may be immediately observed and put aside as philosophically uninteresting. As noted before, the players are able to successfully predict the next state of the system, and thus are able to guide successful action in a causally dynamic world. Indeed, this conclusion could be reached easily without bringing up prediction games at all, since all successful signalling games are successful by virtue of correctly tracking the world by identifying the appropriate action for each observed state. As a result, the players of the prediction game could easily infer that the world must be causal in the minimal sense that it is possible to be reasonably successful in predicting the sequence of events so as to guide action. But there is no open question about whether we are able to describe the world well enough to be able to guide action well. The open question is whether this success indicates anything further about how the world is causally governed, and whether a deeper causal structure may be reliably inferred from the structure of the signalling system. To investigate this, we must look to see what is necessary about any description of the causal laws, and what that says about the five fundamental causal facts noted earlier.

To see which such metaphysical causal facts persist across conventions, and how they do so, consider a different way that the signalling system could have emerged by looking at a different real result (see Table 2).

Sender 1	Signal A	Signal B

Sender 2	Signal X	Signal Y

Receiver	Act 1	Act 2	Act 3	Act 4

State 1	99.4%	0.6%		State 1	99.4%	0.6%		Signal AX	0.06%	98.3%	0.06%	0.06%
State 2	0.6%	99.4%		State 2	0.6%	99.4%		Signal AY	0.6%	0.6%	0.6%	98.3%
State 3	99.4%	0.6%		State 3	0.6%	99.4%		Signal BX	98.3%	0.6%	0.6%	0.6%
State 4	0.6%	99.4%		State 4	99.4%	0.6%		Signal BY	0.5%	0.5%	98.2%	0.7%

Table 2. Probability of each signal, given a state, and each act, given a signal, for a different run of 500,000 plays in a four-state, two-term, two-sender, four-act predictor game, with the causal law followed 90% of the time, and forgetting factor ϕ = 0.90, forgetting period N = 400, and experimental factor ψ = 0.01.

In this signalling system, A refers to the set {1, 3}, B to {2, 4}, X to {1, 4}, and Y to {2, 3}. The causal laws could, again, be put in several equivalent ways:

A and X (state 1)	→	B and Y (state 2)
B and Y (state 2)	→	A and Y (state 3)
A and Y (state 3)	→	B and X (state 4)
B and X (state 4)	→	A and X (state 1)

Alternatively, expressing these laws more in keeping with the players' emergent signalling system:

- A and B always change truth values, regardless of whether X or Y is true;
- X and Y change truth values when A is true, but
- X and Y don't change truth values when B is true.

Finally, the players may simply look at the pattern of terms,

$$A \rightarrow B \rightarrow A \rightarrow B, \quad \text{and} \quad X \rightarrow Y \rightarrow Y \rightarrow X.$$

Contrasting this statement of the laws with that from the earlier signalling system can provide us with some insight into some of what is conventional in these descriptions of the causal laws of the world. We can immediately see that much of the structure has changed, correlating with the conventional elements of that structure noted earlier. Now the evolution of A and B are completely independent of X and Y, where before the evolution of each term depended on the state of the other. The mutual interdependence of the terms, which is an apparently fundamental structural feature of both signalling systems, appears, then, not to be a transcendental fact of the world as it is expressible with the linguistic constraints of the game, and thus seems to be purely conventional. Even the number of laws needed in the second formulation has changed, decreasing from four to three. Any direct inference from these facts to a description of the world, then, will have only the illusion of metaphysical import.

An analysis of all of the causal laws governing the world together, regardless of whether they are understood at the level of the world (one state leading to another), or at the level of the evolved language (one conjunction of terms leading to another), reveals something different, however. In both cases, the same fundamental causal ring topology emerges. Whether a state is understood as simple or complex is certainly conventional, as is whether the parts of a complex state are causally dependent upon each other in any given case, as we saw above. But that states (or conjunctions of states) will repeat, and usually with a regularity of four turns (specifically, they will do so $0.925^4 = 73.2\%$ of the time), and how that regularity governs more specific laws for each transition is identical in both cases (and in all other permutations of a complete signalling system). The more general topology behind the specific causal laws, then, may be said to be non-conventional, regardless of which conventional signalling system emerged.

Furthermore, to understand the complete causal picture through all four turns, the players must understand both terms. Consider the second signalling system outlined above. While it is true that the players may reliably predict the evolution of the first term without reference to the second (A follows B and B follows A, regardless of what is going on with X or Y), they nonetheless must still know the values of both signals (A/B and X/Y) in order to figure out the full next state, since whether X is followed by X or Y is determined by whether A or B is true at the time. And this must be the case. There are no complete signalling systems within which neither term relies on the other; they either both do or only one does. This suggests that the causally efficacious fact is not the value of an individual term, but a conjunction of terms. If the players were metaphysicians, then, and were able to determine this fact, they could end up rejecting the metaphysical dualism of their language in favour of a monism according to which it is the conjunction of two terms that picks out a single causal state. Moreover, this monism would be completely correct, since the true state of the world is monistic, and thus may be a path past the pure conventionalism that Barrett defends. This would then correctly lead to the further conclusion that the linguistic fact of a conjunction should play no role in a true ontology, though this remains contingent upon the players' ability to separate fictional from non-fictional metaphysical propositions, which will be explored in the next section.

It is, finally, worth briefly considering whether these apparently metaphysical facts continue to hold and be expressible in the partial pooling cases. This becomes particularly important as we consider increasingly complex and realistic cases, since the partial pooling rate dramatically increases with complexity. For instance, if we increase the number of terms for each sender to five, then the number of states that they are able to parse is twenty-five. In the twenty-five-state, five-term, two-sender, twenty-five-act game, some partial pooling occurs 89.7% of the time. Nonetheless, with a little fine tuning[4] of the variables for forgetting and experimenting ($N = 2000$, $\phi = 0.96$, $\psi = 0.01$), the players successfully predict the correct outcome around 73% of the time when partial pooling occurs (compared to 4% of the time, were it random chance, and with the next state only

[4] One might object to fine tuning these parameters insofar as any fine tuning will undermine the robustness of the results. This objection may be mitigated in two ways. First, the fine tuning does more to make results apparent and easy to discuss than it drastically affects the results. Second, the fine tuning directly correlates with overall success, without any thresholds, and thus if the parameters were evolvable variables they would plausibly fine tune themselves through the natural process. A robust demonstration of this is beyond the scope of this project.

being the one dictated by the causal law 90.4% of the time, making success above that level impossible), and usually there is only pooling in one or two of the twenty-five combined signal states.

There is nothing particularly revealing about the complexity of the twenty-five-state, five-term, two-sender, twenty-five-act case, however, so for simplicity we will focus on a case of partial pooling for our original four-state, two-term, two-sender, four-act predictor game, which, for our purposes, is essentially identical (Table 3).

Sender 1	Signal A	Signal B	Sender 2	Signal X	Signal Y	Receiver	Act 1	Act 2	Act 3	Act 4
State 1	98.8%	1.2%	State 1	1.2%	98.8%	Signal AX	23.8%	23.8%	23.8%	28.6%
State 2	0.6%	99.4%	State 2	99.4%	0.6%	Signal AY	0.6%	46.7%	0.6%	52.1%
State 3	98.9%	1.1%	State 3	1.3%	98.7%	Signal BX	0.6%	98.4%	0.6%	0.6%
State 4	0.6%	99.4%	State 4	0.6%	99.4%	Signal BY	98.4%	0.6%	0.6%	0.6%

Table 3. Probability of each signal, given a state, and each act, given a signal, for a run of 500,000 plays in a four-state, two-term, two-sender, four-act predictor game, with the causal law followed 90% of the time, and forgetting factor $\phi = 0.96$, $N = 2000$, and experimental factor $\psi = 0.01$.

Here, Signal A corresponds to {1, 3}, Signal B corresponds to {2, 4}, Signal X corresponds to {2}, Signal Y corresponds to {1, 3, 4}. The partial pooling causes Signal AY to correspond both to State 1 and to State 3. Given that, the causal laws change a bit:

A and Y (state 1)	→	B and X (state 2)	or	B and Y (state 4)

B and X (state 2)	→	A and Y (state 3)
B and Y (state 4)	→	A and Y (state 1)

Our second statement of the laws then becomes:

- A, B, and X always change truth values, regardless of what else is true.
- Y sometimes changes its truth value when A is true, but never when B is true.

Finally, the pattern of terms would look like this:

A → B → A → B,   and   Y → X → Y → Y.

What of our earlier conclusions? Do the metaphysical facts we identified earlier continue to hold true? Three related metaphysical facts were determined above as partial representations of the basic causal structure and laws as they exist from outside the game: first, the ring topology of the causal laws; second, the interdependence of the two terms; third, state monism. The dependence of the two terms on each other remains, though it fails to be completely predictive, since they would also now need to know the previous state in order to determine whether Y is changing to X or not. Given the need to consider the whole state, the broader four-step ring topology also remains: AY → BX → AY → BY. There is a smaller repetition of AY every other turn, but, nonetheless, the four-step ring remains for the system as a whole. Similarly, since knowledge of both signals is needed to be as successfully predictive as possible, state monism is again a metaphysical truth.

This is only half of what is needed to avoid a broadly sweeping scepticism that would block inference to the conditional claims discussed above (if there are metaphysical facts, then they are like this). There must be metaphysical facts that transcend accidental facts about language, several of which we've now identified and shown to be (somewhat) consistent upon different results of the game, but those facts must also be discoverable and separable from the merely conventional structure of the language and fictional interpretations. That we are in a position to do so from our perspective outside the game is, obviously, not a reason to think that the players could, and it is what is discoverable by the players that is philosophically interesting and that corresponds to the possibility of legitimate metaphysical inquiry.

6 Learned and Invented Metaphysics

The discussion above about the emergent metaphysical facts present in the game—the causal ring topology, the interdependence of the two terms for the successful prediction of the next state, and state monism—relied on an analysis of the past history of successful predictions and states. At first glance, this would seem to keep these facts hidden from the players, since they only have the ability to use information about the current state or signal to guide action. However, just because they cannot look at specific historical events when deciding action does not mean that the players have no record of past history.

Rather, that history is recorded through the learning algorithm, and thus each decision does, in a way, appeal to the results of past decisions. Therefore, the players do seem to have access to some of the resources necessary to be able to describe the metaphysical facts we have identified.

Two points of caution are needed with this conclusion, however. First, being able to describe true metaphysical facts does not in any way contradict or limit Barrett's scepticism. Barrett has proven that there may be nothing but conventions, regardless of how successful the language is. The goal, rather, is to determine whether it is possible to determine how the world would be if there were some deeper structure governing the world, and this question remains open. Second, merely having the resources to be capable of stating a true metaphysical claim doesn't prove much. I have the linguistic ability to state tomorrow's winning lottery numbers, since I have the ability to state any finitely expressible string of numbers I'd like. What would be of particular interest to the present discussion would be if the players could differentiate apparently metaphysical claims that are purely conventional facts of their language or are outright fictions from those that are true or at least would be true if there were some fact to correspond to. For instance, we've seen that, in the evolved signalling system, there are multiple kinds of state—corresponding to A/B and X/Y—though as is obvious from our perspective these are merely conventional divisions into fictional kinds. The same is true for any of the three different ways of describing the causal laws in a given signalling system described above. One could also imagine a complete fiction, such as a statement that there is an undetectable hidden variable that deterministically fixes the evolution of the world from one state to the next. Can the players separate these claims from the three facts discussed above?

To begin, we need to rule out two apparent paths to separating the conventional and fictional from truly metaphysical claims discussed above. They obviously cannot directly compare their claims to the actual state of the world and they cannot compare one evolved language to another. Just as we cannot do much more than guess at how science would have been had it evolved differently, sameness between evolved signalling systems is an epistemic non-starter.

To more easily see the precise problem that the players face, consider the results of the partial pooling signalling system discussed above. In that case, state AY sometimes is followed by BX and sometimes by BY. Nonetheless, there is the same four step probabilistic loop as in the other signalling systems, excepting, of course, that AY appears twice in the loop. How might this be interpreted by the players? On the one hand, it is possible to interpret this basically correctly and determine that the players have failed to divide nature as precisely as possible, that two different states are being picked out by AY. This still leaves somewhat open to the players the question of whether, once things were divided properly, the states should be understood as compound or simple with a compound name, but this is promising insofar as that is made a ‘scientific’ question. There are, however, several other interpretations that seem to be available.

Looking more closely at the details of the four state loop, one could expand the second formulation of the causal laws as follows:

- A, B, and X always change truth values, regardless of what else is true.
- Y changes its truth value when A is true and Y was true the previous round.
- Y doesn't change its truth value when B is true or when Y wasn't true the previous round.

As before, this may suggest that Y isn't doing a very good job of picking out discrete states of the world, but it could just as well be taken to suggest that Y's causal relevance extends beyond the current state, a completely fictional (though predictively effective) conclusion, since from a god's eye view only the current state actually matters.

It is also possible for the players to interpret nothing further into the situation. There is no reason why it must be the case that each state brings about the next state with an identical 92.5% probability. It is perfectly coherent to hold that there is a deep indeterminism that happens to be split between two possible results when AY is the case, so that BX and BY each have a 46.25% probability of coming next. This could correctly suggest both the interdependence of the two terms and state monism again, but speaks against an accurate understanding of the basic causal structure.

Finally, combining parts of each of the previous three interpretations, the players might conclude that there is a missing variable at work that is doing the work of 'deciding' whether BX or BY is the more likely state to follow from AY in a particular play of the game. This isn't far from correct, of course, except insofar as there is no particular reason for the players to conclude that there is some hidden factor that just explains this one uncertainty and not the rest. Once the possibility of a hidden variable has been introduced, it is a short step to thinking that some hidden variable in fact explains all of the apparent indeterminism in the evolution from one state to the next, and we can see that there is nothing particularly special about the partial pooling case for wrestling with indeterminism. Depending on what hidden variables are proposed, it may or may not be the case that state monism or term interdependence emerges.

While it is true that each of these interpretive theories do contain at least the possibility of some version of some of the metaphysical facts that were discussed above, none of the theories unambiguously have all three facts and they all have some further fictional theses that seem to be as supported as the nonfictional ones. This suggests, then, that when partial pooling occurs, there is no reliable way for the players to describe how the world would be if their descriptive scheme were basically correct and if there were some deeper structure that can be described with the evolved proto-language (and some liberties with meta-analysis).

One might be forgiven for thinking that this is a particular shortcoming with partial pooling signalling systems. After all, when partial pooling doesn't occur, all three facts are clearly in evidence and no obvious fictions were identified other than those, such as state dualism, which could be put aside with further analysis. This is erroneous for a few reasons. First, two of the erroneous interpretations carry over from the partial pooling case to the full signalling system case: the possibility of hidden variables and the two-step causal influence. The possibility of the first of these in the non-partial pooling case was noted above, with the hidden variable explaining all indeterminacy, rather than the unequal indeterminacy. To see how the two-step causal influence may translate to the non-pooled signalling system, consider again the second signalling system discussed above. The original interpretation provided was that whether X is followed by Y or by another X is determined by whether it is conjoined with A or with B, which was how signal interdependence and state monism were deduced. An equally legitimate (though fictional) interpretation is, rather, that whether X is followed by Y is determined by whether X or Y was true in the previous step.

The problem here isn't really that that is incorrect, but that it would be just as plausible if it were correct or incorrect, and is only slightly more speculative than, for instance, state monism. More generally, the 'world' could have been structured in such a way that each of the fictions we've encountered is actually correct, without any change to the evolutionary process or product. Furthermore, it could be the case that the world is best described with partial pooling. This would be the case, for instance, if there were a fork in the causal system: state one usually leads to either state two or state four, which in turn both usually lead to state one. In this case, the partial pooling signalling system would be as successful as possible (though there are a large number of other equally successful signalling systems), and would do a better job of describing the actual causal system than would a non-pooling signalling system. Furthermore, there is no standpoint available for the players that can reveal the one system to be error and the other correct.

One might suspect that, if the players could become much more sophisticated, then they may eventually be able to discover some of these differences, even if they are invisible from the players' perspective so far as this game allows. This might be possible, for instance, if 'researchers' within a more complex game could experiment to isolate causes and directly test pluralism and monism. For this to make progress against the scepticism defended here, however, requires that the world is completely empirically available to the players. If there are, for instance, some real empirically hidden variables, then it remains impossible for the players to distinguish between emergent metaphysical truths and fictions or conventions. It is only possible for greater linguistic sophistication to distinguish metaphysical truths from conventions or fictions if metaphysical distinctions are empirically discoverable, the exact question at issue. Indeed, it must be empirically discoverable that metaphysical distinctions are empirically discoverable, which is what Barrett's original argument refuted. Without that knowledge being given, even if everything that emerges from the players' linguistic structure were metaphysically true, the players would be incapable of knowing that none of their emergent metaphysics is fictional.

Combining these results with those of Barrett's argument warrants a more robust metaphysical scepticism. Barrett has shown that metaphysical theories derived from linguistic structures do not require metaphysical truths, even when we have a language that is as empirically successful as the world permits. The present argument has explored the state of things when there are metaphysical facts present—a fact that is, of course, impossible to determine one way or another, according to Barrett. We have now seen that it is possible to both describe some of these metaphysical facts with the evolved signalling system, and in most signalling systems and interpretations at least some of them are deducible from the structure of the language and its success in describing the world. However, this is not always the case, and almost inevitably some fictions will be just as deducible and the players lack the resources necessary to be able to separate truth from fiction in their metaphysical theories. Thus, in addition to being unable to determine whether there is any metaphysical structure to discover, they also cannot reliably reach conditional conclusions along the lines of, 'if there are true metaphysical facts, then the success and structure of our language says that they must be this way rather than that way'.

Gordon Michael Purves
Philosophy, Theology, and Religious Studies
Sacred Heart University College of Arts and Sciences
Fairfield, CI, USA
purvesg@sacredheart.edu

References

Alexander, J. M. [2014]: 'Learning to Signal in a Dynamic World', *British Journal for the Philosophy of Science*, **65**, pp. 797–820.

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**, pp. 373–90.

Barrett, J. [unpublished]: 'Numerical Simulations of the Lewis Signaling Game: Learning Strategies, Pooling Equilibria, and the Evolution of Grammar', available at <escholarship.org/uc/item/5xr0b0vp>.

Barrett, J. [2007]: 'Dynamic Partitioning and the Conventionality of Kinds', *Philosophy of Science*, **74**, pp. 527–46.

Barrett, J. [2014a]: 'Rule-Following and the Evolution of Basic Concepts', *Philosophy of Science*, **81**, pp. 829–39.

Barrett, J. [2014b]: 'On the Coevolution of Theory and Language and the Nature of Successful Inquiry', *Erkenntnis*, **79**, pp. 821–34.

Barrett, J. and Zollman, K. J. [2009]: 'The Role of Forgetting in the Evolution and Learning of Language', *Journal of Experimental and Theoretical Artificial Intelligence*, **21**, pp. 293–309.

Barrett, J., Dickson, M. and Purves, G. [unpublished]: 'Evolutionary Prediction Games', available at <philsci-archive.pitt.edu/10042/>

Dickson, M. [2013]: 'Theory from Chaos', *Episteme*, **10**, pp. 465–78.

Goodman, N. [1983]: *Fact, Fiction, and Forecast,* Cambridge, MA: Harvard University Press.

Huttegger, S. [2007a]: 'Evolution and the Explanation of Meaning', *Philosophy of Science*, **74**, pp. 1–27.

Huttegger, S. [2007b]: 'Evolutionary Explanations of Indicatives and Imperatives', *Erkenntnis*, **66**, pp. 409–36.

Huttegger, S. M., Skyrms, B., Smead, R. and Zollman, K. J. [2010]: 'Evolutionary Dynamics of Lewis Signaling Games: Signaling Systems vs. Partial Pooling', *Synthese*, **192**, pp. 177–91.

Larsson, M. [2015]: 'Tool-Use-Associated Sound in the Evolution of Language', *Animal Cognition*, **18**, pp. 993–1005.

Lewis, D. [1969, 2002]: *Convention: A Philosophical Study,* Oxford: Blackwell.

O'Connor, C. [forthcoming]: 'Games and Kinds', *British Journal for the Philosophy of Science*.

Roth, A. and Erev, I. [1995]: 'Learning in Extensive-Form Games: Experimental Data and Simple Dynamics Models in the Intermediate Term', *Games and Economic Behavior*, **8**, pp. 164–212.

Seyfarth, R. M., Cheney, D. L. and Marler, P. [1980]: 'Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication', *Science*, **210**, pp. 801–3.

Skryms, B. [2008]: 'Signals', *Philosophy of Science*, **75**, pp. 489–500.

Skyrms, B. [2009]: 'Evolution of Signalling Systems with Multiple Senders and Receivers', *Philosophical Transactions of the Royal Society B*, **364**, pp. 771–9.

Skyrms, B. [2010]: *Signals: Evolution, Learning, and Information*, Oxford: Oxford University Press.