ON THE EVOLUTION OF TRUTH

JEFFREY A. BARRETT

Abstract. This paper is concerned with how a simple metalanguage might coevolve with a simple descriptive base language in the context of interacting Skyrms-Lewis signaling games [16, 19, 7]. We will first consider a metagame that evolves to track the successful and unsuccessful use of a coevolving base language, then we will consider a metagame that evolves to track the truth of expressions in a coevolving base language. We will see how a metagame that tracks truth provides an endogenous way to break the symmetry between indicative and imperative interpretations of the base language. Finally, we will consider how composite signaling games provide a way to characterize alternative pragmatic notions of truth.

1. Introduction

Language may describe the world directly, but language may also describe language. An expression might describe the number of letters in a word or more subtle facts such as whether a particular utterance led to successful action or whether a particular statement provides a faithful description of the world. Here we are concerned with how a metalanguage might coevolve with the language it describes.

Skyrms-Lewis signaling games illustrate how it is possible for agents with limited dispositional resources to evolve successful signaling systems and languages with simple grammars as the agents interact with the world and each other.[1] Signaling games may also evolve to interact with each other to form more complex games.[2] Here we will consider the composition of a base game and a metagame that takes states of the base game as input, and we consider how such a composite game may coevolve a simple descriptive language and a metalanguage that is descriptive of the base language and its use.

The first metagame we consider coevolves to indicate the success and failure of the base-game agents as they evolve and use a very simple language. In the second metagame, the sender attends to the coevolving customary use of expressions in the base game. In those cases where the base game evolves a successful signaling system, this metagame evolves to track whether the expressions of the base-game

Date: April 5, 2015.

[1]See David Lewis's [16] characterization of signaling games. See Barrett [4] for an example of the evolution of a simple grammar in a two-sender signaling game and Skyrms [19] for an overview of recent signaling games.

[2]See Barrett and Skyrms [7] for a discussion of how simpler signaling games may combine to form more complex games by cue-reading, template transfer, and modular composition.

sender are true. Insofar as the metagame evolves to track the truth of the base-game sender's expressions, the model provides a sense in which the base language might be understood to have evolved propositional content. Such composite signaling games also provide a way to characterize alternative pragmatic notions of truth.

2. SUCCESS AND FAILURE

Consider two signaling games: a base game that takes states of nature as input and a metagame that takes states of the base game as input.

The base game is a $4 \times 4 \times 4$ signaling game with unbiased nature where the agents learn by simple reinforcement.[3] In this game there are four possible states of nature, each requiring a different receiver action for success. Nature chooses a state at random. The sender has four urns, each corresponding to one of the possible states of nature and each beginning with one ball corresponding to each of the four possible signal types. The sender observes the state of nature, draws a ball at random from the corresponding urn, then sends the signal indicated by that ball. The receiver, who has no direct access to nature, has four urns, each corresponding to one of the possible signals and each beginning with one ball corresponding of each of the four possible actions. The receiver observes the signal, draws a ball from the corresponding urn, then performs the corresponding action. If the action matches the state of nature, it is successful, and the sender and receiver each return the ball they drew to the urn from which it was drawn and add another ball of the same type to that urn. If the action does not match the state of nature, they just return the balls to the urns from which they were drawn.

The signals are initially meaningless and are used randomly. Consequently, the receiver's actions are typically unsuccessful. As the agents learn, however, the signals evolve meanings that communicate information that serves to coordinate the actions of the receiver to the states of nature.[4] On simulation, the receiver exhibits a cumulative success rate of better than 0.95 about 0.75 of the time, with 1000 runs of 1×10^6 plays per run. The rest of the time the game gets stuck in a suboptimal partial pooling equilibrium that exhibits a cumulative success rate of about 0.75.[5]

---

[3]See Herrnstein [12] for a discussion of simple reinforcement learning and Roth and Erev [18, 9] and Huttegger, Skyrms, Tarrès, Wagner [15] for discussions of more subtle forms of reinforcement learning and other options. In the case of the simplest varieties of reinforcement learning, one might imagine the agents learning by adjusting the contents of urns on the basis of their experience as described here.

[4]See Skyrms [19] for a discussion of the precise sense in which the evolved signals communicate information.

[5]See Barrett [5] for further details regarding the behavior of this and closely related signaling games.

Note that when the base game evolves to match states of nature to successful actions there is a symmetry between indicative and imperative interpretations of the signals. In particular, they might be interpreted as indicatives where the sender reports the state of nature or as imperatives where the sender tells the base-game receiver what to do. In either case, however, when a successful signaling system evolves, the signals transmit information regarding the state of nature, which is reflected in the uniform success of the receiver's actions.[6]

The metagame is a $2\times2\times2$ signaling game where a second pair of agents learn by simple reinforcement.[7] The metagame sender takes the success and failure of the agents in a particular play of the base game as input. It is the play of the base game, then, that provides the states of nature of the metagame (Figure 1).

The metagame agents also learn by simple reinforcement. The metagame sender has two urns, one corresponding to success and one corresponding to failure in the base game. Each of these urns begins with one ball of each of the two possible signal types. The metagame sender observes whether the current play of the base game is successful. This might involve observing the state of nature and the action in the base game then checking whether they match, observing whether the agents in the base game reinforce, or just directly observing whether the base-game receiver's action produces a result that in fact indicates success given the second-order dispositions of the agent to reinforce. The metagame sender then draws a ball at random from the corresponding urn and sends the signal indicated on that ball. The metagame receiver has two urns, each corresponding to one of the two possible signals and each beginning with one ball corresponding of each of the two possible actions. One metagame action type is successful if and only if the current play of the base game was successful, the other is successful otherwise. The metagame receiver observes the signal, draws a ball from the corresponding urn, then performs the corresponding action. If his action was successful, then the metagame sender and receiver each return the ball they drew to the urn from which it was drawn and add another ball of the same type to that urn. If it was unsuccessful, then they just return their balls to the urns from which they were drawn.

Here the dispositions of the metagame agents coevolve with the dispositions of the base-game agents. The metagame sender begins by randomly sending signals

[6]Again, see Skyrms [19] for a characterization of how one might understand information transfer in the context of signaling games.

[7]The composite system might be taken to model agents observing the evolving language use of other other agents or agents observing their own evolving language use. See [7] for a discussion of such composite systems in nature and how such complex games might self-assemble from simpler dispositions by way of evolutionary processes. Such models explain *how it is possible* for relatively sophisticated linguistic competences to evolve in the context of modest dispositional resources. They also support the view that one might individuate alternative pragmatic notions, represented here in the coevolving metagame language, by the games that evolves them.

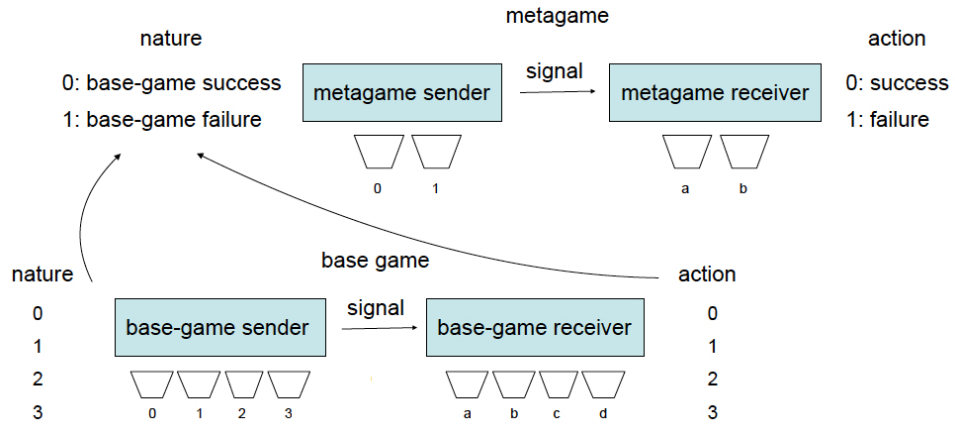

FIGURE 1. Composite success-failure game

and the metagame receiver begins by randomly acting on the signals. But they learn from their observations of the evolving dispositions of the base-game agents. On simulation, the metagame receiver exhibits a cumulative success rate of better than 0.95 on better than 0.99 of the runs of the model, with 1000 runs of 1×10^6 plays per run. An action is successful here if and only if it correctly matches whether the base-game agents were in fact successful on the current play of their game.

Interpreting the success of the metagame, however, requires some care. It is known that a 2×2×2 signaling game where the agents learn by simple reinforcement will converge to a successful signaling system if nature is unbiased, but it is also known that it may get stuck in a suboptimal partial pooling equilibrium if nature is biased.[8] Nature for the metagame is the evolving success and failure of the base game, which is strongly biased toward success over time. Indeed, as we have seen, the base game typically evolves to exhibit a nearly perfect cumulate success rate. When it does, the metagame might exhibit a similarly high cumulative success rate if the receiver always acted as if the base game were successful regardless of the signal he gets. Those successful dispositions would be reinforced in the metagame, and both terms would evolve to be associated with success. In this case, the metagame would be successful, but it would not coevolve the expressive resources to represent both success and failure in the base game.[9]

[8]See Argiento, Pemantle, Skyrms, and Volkov [1] for a proof of the first point, and see Hofbauer and Huttegger [14] for a proof of the second point in the context of a population model under replicator dynamics.

[9]Note that we are not assigning indicative content to the evolved signals of the base game or the metagame in this case. Given the symmetry of the two games, the expressions in either game might be interpreted as indicatives or imperatives. The signals in the metagame coevolve to communicate information concerning the success or failure of plays of the base game to the metagame receiver and this information is reflected in his successful actions given the state of

But that is not what happens here. Rather, since the base game is significantly more complicated than the metagame, and hence evolves more slowly, by the time the base-game agents have evolved meaningful signals, the metagame agents have had plenty of time to evolve signals that distinguish between successful and unsuccessful plays of the base game. More specifically, on simulation, the metagame agents evolve signals that distinguish sharply between success and failure in the base game better than 0.98 of the time, with 1000 runs of 1×10^6 plays per run.[10] In short, on the present model, the metagame agents have the chance to observe enough unsuccessful signaling to learn to make the distinction between successful and unsuccessful signaling in the base game.

The base-game signals, then, typically evolve to correspond to the four states of nature that they individuate and the actions that successfully match those states. And the metagame signals typically evolve to communicate information concerning whether the base-game agents were successful on a particular play of their game.

Note that while an incorrect action in the base game may be the result of the sender using the wrong signal or the receiver doing the wrong thing when he gets the signal or some combination of both, the metagame does not evolve the expressive resources to indicate which agent was responsible for an unsuccessful base-game action. Rather, it just evolves to track whether particular plays of the base game were successful. And since it only tracks whether the base-game agents were successful or unsuccessful together, it does not brake the symmetry between indicative and imperative interpretations of the base-game signals. These signals may equally well be interpreted as indicatives where the base-game sender reports the state of nature or as imperatives where the she tells the base-game receiver what to do.

A metagame might, however, coevolve to track whether the base-game sender used the *right* signal on a particular play of the game given the current state of nature and how the base-game conventions have in fact evolved. When the base game evolves a successful signaling system, such a metagame game might be understood as tracking whether the base-game sender's signal represents the current state of nature. As such, the signals in the metagame would communicate information concerning the truth of the base-game sender's signal. And, insofar as

---

nature. We will consider how the symmetry between indicative and imperative interpretations of the signals might be broken in the next section.

[10]More specifically, the magnitude of the dispositions that individuate signals nearly always differ by better than two orders of magnitude, and typically significantly more. The metagame agents do yet better when they learn by way of a faster dynamics like reinforcement with punishment or forgetting. See Roth and Erev [18, 9] and Barrett and Zollman [8] for descriptions of learning dynamics that are both faster and more reliable in the context of such games.

one understands the metagame as tracking the truth of the base-game signals, they might be taken as indicatives representing the states of nature.[11]

3. TRUE AND FALSE

Metagame signals that track whether the base-game sender used the customary signal given the current state of nature may coevolve when the metagame sender attends to how the base-game sender's use of the base-game signals evolve. If so, and if the base-game signals come to be associated with particular states of nature, the metagame signals might be interpreted as tracking the truth of the base-game sender's descriptions of nature.

Consider the same $4\times4\times4$ base game and a new $2\times2\times2$ metagame. The new metagame agents learn by simple reinforcement, but here the metagame sender observes whether the base-game sender sent the signal in the current play that she has used most often to this point when she has observed the current state of nature. And the metagame receiver has two new actions. One is successful if and only if the base-game signal was the most often used given the current state of nature and the other is successful otherwise. The metagame agents learn by simple reinforcement on the success of the metagame receiver (Figure 2).

On simulation, the metagame agents typically evolve signals that distinguish well between the base-game sender sending the signal that she has sent most of the time in the current state of nature and her not doing so. More specifically, they evolve sharply distinguishing signals approximately 0.78 of the time on 1000 runs of 1×10^6 plays per run.[12] When successful, the metagame sender's signals coevolve to represent whether or not the base-game sender used her current signal in the customary way.

---

[11] See Harms [10, 11] and Millikan [17] for discussions of the propositional content of evolved languages and Huttegger [13] and Zollman [20] for discussions of alternative approaches for breaking the symmetry between indicative and imperative interpretations of the evolved signals in Skyrms-Lewis signaling games. In short, Huttegger's approach is to introduce deliberation as an additional primitive option in the game then individuate interpretations of the evolved signals by whether the sender or receiver choose to deliberate, and Zollman's is to introduce a second receiver then consider a game where the sender might assert to both receivers or direct each separately. The present proposal, rather, is to break the symmetry by allowing for the evolution of a metalanguage that tracks the sender's use of the base language given the current state of nature and how the base language has evolved. It is likely that the symmetry between indicatives and imperative is broken in multiple, context-dependent ways in the evolution and use of natural languages.

[12] When they fail to do so, both metagame signals evolve to indicate that the base-game sender failed to send the signal most often used in the current situation. Since the metagame uses simple reinforcement learning, one would expect this suboptimal equilibrium from time to time on runs where the base-game sender is slow to converge to a set of stable, surefire dispositions. One would also expect such suboptimal behavior to be less likely if the metagame agents were to learn by a form of reinforcement with punishment or forgetting. See Roth and Erev [18, 9] and Barrett and Zollman [8] for descriptions of such learning dynamics.

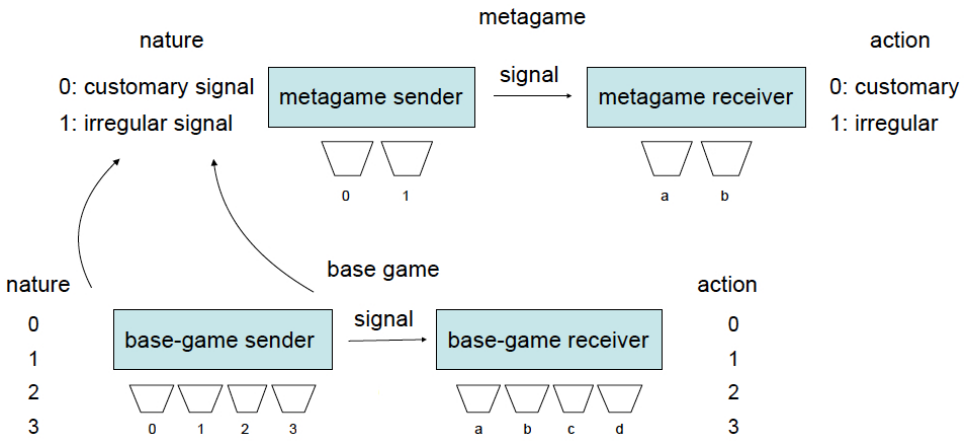

FIGURE 2. Composite true-false game

In consequence, if the metagame signals coevolve to communicate whether the base-game sender sent the customary signal given the current state of nature and if the base game evolves a successful signaling system that matches states of nature to successful actions, then the metagame signals will also successfully communicate whether or not the base-game sender's signal correctly represents the current state of nature given the evolved base-game conventions. Under these conditions, then, the metagame will coevolve to track whether the base-game sender's signal is *correct* given the current state of nature and the evolved base-game conventions. And it will do so independently of the success or failure of the base-game receiver's action on a particular signal. That is, when the base-game evolves a successful signaling system, the metagame typically coevolves to track the *truth* of the base-game sender's signals.

Insofar as they may be taken as true or false from the perceptive of the metagame, the base-game sender's signals, then, are good candidates for indicatives that either succeed of fail to correctly describe of nature. This provides an endogenous means of breaking the interpretational symmetry between indicative and imperative interpretations of the base-game signals.[13]

The present model illustrates how it is possible, with modest dispositional resources, to coevolve a simple descriptive base language and a metalanguage that

[13]Note that while there is nothing here that breaks the symmetry of the interpretation of metagame signals, however one understands the metagame signals, when the base-game evolves a signaling system, the metagame signals communicate information, in the sense characterized by Skyrms [19], concerning whether the base-game sender's signal faithfully represents the current state of nature.

tracks the truth of evolved base-game expressions.[14] More generally, one might expect a more sophisticated base language to evolve in a more subtle base game and a more sophisticated metagame to track more subtle aspects of how expressions in that base language may relate to the history of successful use and the nature of the world. The compositions of such games would provide correspondingly richer languages and hence richer evolutionary accounts of truth.[15]

4. Discussion

In the first metagame, the agents learn to distinguish between successful and unsuccessful language use in the base game. In this case, the evolution of a meaningful metalanguage does not require the base game to evolve a successful signaling system. Indeed, the faster the base game evolves toward optimal signaling the more difficult it is for the metagame agents to evolve signals that track success and failure in the base game.

In the second metagame, the agents coevolve a metalanguage that tracks the truth of the base-game expressions by attending to whether the base-game sender uses her signals in the evolved customary way. Here the evolution of a truth predicate in the metagame requires that the base-game agents in fact evolve a successful signaling system.

Insofar as the second metagame can be understood as tracking whether the base-game expressions provide true descriptions of nature, the base-game expressions can, in turn, be understood as having propositional content. Namely, they communicate the state of nature that currently obtains. As in the first metagame, the resulting metalanguage is weakly normative in what it evolves to communicate; in this case, the coevolving conventional use of expressions in the base language.

It is a significant feature the present model that the metagame evolves to track the truth of the base-game expressions even as these expressions themselves evolve meanings. Before there is a meaningful base-game language, there are no expressions that might be true or false. And, before the base-game language evolves to allow for successful action, there is nothing to tie the customary use of the base-game expressions with what is in fact true.

The composite game, then, provides a simple pragmatic model of truth. On a pragmatic notion of truth, a language comes to allow for true descriptions of nature only as it comes to allow for successful coordinated action. On the present model, it

[14]The evolved distinction between true and false is available to represent possible failures in future plays of the base-game agents due to a broken or deceptive sender. Such use would further reinforce the metagame distinction.

[15]A metagame like the one described here that is associated with the base game described in Barrett [3] might, for example, coevolve to track a primitive sort of truth for very basic arithmetic statements.

is only when the base game agents evolve a language that communicates information regarding nature that allows for successful coordinated action that the metagame coevolves a language that communicates the truth of the base-game expressions.[16]

One would expect more subtle evolutionary models to pick out alternative, richer, varieties of pragmatic truth.[17] The thought here is that one might individuate alternative pragmatic notions of truth by the games that evolve them. On this view, the pragmatic notion of truth illustrated in the present model would be among the simplest that connect successful use to one's coevolved representations.
[18]

[16]See Barrett [2, 6] for discussions of how faithful description might coevolve with successful inquiry.
[17]Kevin Zollman, for example, suggested a natural extension of the present metagame where the metagame receiver uses the metagame signal to decide whether to use the base-game signal as a basis for action.
[18]I would like to thank Andrew Bollhagen, Brian Skyrms, Simon Huttegger, Cailin O'Connor, and Kevin Zollman for helpful discussions. I would also like to thank two anonymous reviewers for their very helpful comments on an early version of this paper.

References

[1] Argiento, Raffaele, Robin Pemantle, Brian Skyrms and Stas Volkov (2009) "Learning to Signal: Analysis of a Micro-Level Reinforcement Model," *Stochastic Processes and Their Applications* 119(2): 373–390.

[2] Barrett, J. A. (2014) "On the Coevolution of Theory and Language and the Nature of Successful Inquiry," *Erkenntnis* 79(4): 821–834.

[3] Barrett, J. A. (2013a) "On the Coevolution of Basic Arithmetic Language and Knowledge" *Erkenntnis* 78(5): 1025–1036

[4] Barrett, J. A. (2007a) "Dynamic Partitioning and the Conventionality of Kinds," *Philosophy of Science* 74: 527–546.

[5] Barrett, J. A. (2006) "Numerical Simulations of the Lewis Signaling Game: Learning Strategies, Pooling Equilibria, and the Evolution of Grammar," *Institute for Mathematical Behavioral Sciences* Paper 54. http://repositories.cdlib.org/imbs/54.

[6] Barrett, J. A. (2001) "Toward a Pragmatic Account of Scientific Knowledge," *PhilSci Archive*. uri: http://philsci-archive.pitt.edu/498/ accessed 20 August 2014.

[7] Barrett, J. A. and B. Skyrms (2015) "Self-Assembling Games," forthcoming in *The British Journal for the Philosophy of Science*.

[8] Barrett, J. A. and Kevin Zollman (2009) "The Role of Forgetting in the Evolution and Learning of Language," *Journal of Experimental and Theoretical Artificial Intelligence* 21(4): 293-309.

[9] Erev, I. and A. E. Roth (1998) "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria" *American Economic Review* 88: 848–81.

[10] Harms, W. F. (2004a) *Information and Meaning in Evolutionary Processes.* Cambridge: Cambridge University Press.

[11] Harms, W. F. (2004b) "Primative Content, Translation, and the Emergence of Meaning in Animal Communication," in D. Kimbrough O. and U. Griebel (eds.) *Evolution of Communication Systems: A Comparative Approach.* Cambridge: MIT Press.

[12] Herrnstein, R. J. (1970) "On the Law of Effect," *Journal of the Experimental Analysis of Behavior* 13: 243–266.

[13] Huttegger, Simon (2007) "Evolutionary Explanations of Indicatives and Imperatives. Erkenntnis 66, 2007, 409-436.

[14] Hofbauer, Josef and Simon Huttegger (2008) "Feasibility of Communication in Binary Signaling Games," *Journal of Theoretical Biology* 254(4): 843–849.

[15] Huttegger, Simon, Brian Skyrms, Pierre Tarrès, and Elliott Wagner (2014) "Some Dynamics of Signaling Games," *Proceedings of the National Academy of Sciences* 111(S3): 10873–10880.

[16] Lewis, David (1969) *Convention.* Cambridge, MA: Harvard University Press.

[17] Millikan, R. G. (2005) *Language: A Biological Model.* Oxford: Oxford University Press.

[18] Roth, A. E. and I. Erev (1995) "Learning in Extensive Form Games: Experimental Data and Simple Dynamical Models in the Immediate Term," *Games and Economic Behavior* 8:164–212.

[19] Skyrms, Brian (2010) *Signals Evolution, Learning, & Information.* New York: Oxford University Press.

[20] Zollman, K. J. S. (2011) "Separating Directives and Assertions Using Simple Signaling Games," *The Journal of Philosophy* 63(3): 158–169.

UC Irvine; Irvine, CA 92697, USA
E-mail address: j.barrett@uci.edu