Author Names: Hannah Rubin, Cailin O'Connor, and Justin Bruner

Title: Experimental Economics for Philosophers

Abstract: Recently, game theory and evolutionary game theory – mathematical frameworks from economics and biology designed to model and explain interactive behavior – have proved fruitful tools for philosophers in areas such as ethics, philosophy of language, social epistemology, and political philosophy. This methodological osmosis is part of a trend where philosophers have blurred disciplinary lines to import the best epistemic tools available. In this vein, experimental philosophers have drawn on practices from the social sciences, and especially from psychology, to expand philosophy's grasp on issues from morality to consciousness. We argue that the recent prevalence of formal work on human interaction in philosophy opens the door for new methods in experimental philosophy. In particular, we discuss methods from experimental economics, focusing on a small literature we have been developing investigating signaling and communication in humans. We describe results from a novel experiment showing how environmental structure can shape signaling behavior.

Keywords: Experimental Economics, Induced Valuation, Game Theory, Evolutionary Game Theory, Sim-Max Games, Communication, Convention, Philosophy of Language

1. Introduction

Over the last twenty years or so, game theory and evolutionary game theory - mathematical frameworks from economics and biology designed to model and explain interactive behavior - have proved fruitful tools for philosophers. Ethics, philosophy of language, philosophy of cognition and mind, social epistemology, philosophy of biology and social science, and social and political philosophy, for example, all focus on questions related to human interaction, meaning that game theory and evolutionary game theory have been useful in illuminating problems of traditional interest in these fields.

This methodological osmosis is part of a larger trend where philosophers have blurred disciplinary lines in order to use the best epistemic tools available when tackling the questions that interest them. In this vein, experimental philosophers have drawn on practices from the social sciences, and especially from psychology, to expand philosophy's grasp on issues from morality to epistemology to consciousness.

In this paper, we argue that the recent prevalence of formal work on human interaction in philosophy opens the door for new methods in experimental philosophy. In particular, we discuss methods from experimental economics, focusing on studies of strategic behavior, to show how these methods can supplement, extend, and deepen philosophical inquiry. This branch of experimentation emphasizes induced valuation - the idea that if we want to understand strategic behavior in humans, we have to create a situation which mimics the strategic structure of the world. In other words, we have to allow people to make real choices that will impact actual outcomes that they value, as opposed to, say, reporting what choices they would make in such a scenario. The experimental framework also uses minimal framing, on the assumption that we are looking for general behavioral patterns. This contrasts with some commonly used methods in experimental philosophy that emphasize responses to specific cases and speculation on the counterfactual behavior of the subject.

We will ground our discussion of these methods in a small literature we have been part of developing that uses experimental economics to investigate signaling, language, and communication in humans. In particular, we will describe two studies we have recently completed. The first asks: under what conditions does common interest communication arise in small experimental groups? The second asks: can partially honest patterns of communication emerge between humans? We will also present a novel study on the emergence in communication in humans. We consider how the structure of the world that people encounter impacts the languages that emerge. In particular, we ask: do similarity structures ease the development of conventional terms, especially in complex worlds?

As we will argue, these studies are important complements to the theoretical work that inspired them. They lend credence to evolutionary game theoretic predictions, both in the specific cases, but also as a general tool for predicting human communicatory behavior. In this way, they play a double epistemic role, telling us something about human behavior as well as about our other methods for understanding human behavior. In sum, we argue that these experimental methods have much to offer experimental philosophy, for extending and improving existing game theoretic explorations in philosophy, but also for any inquiry into the nature of strategic interaction - cooperation, altruism, communication, social coordination, social learning, etc. - in humans.

In section 2 we will describe the methods we import from experimental economics. In section 3 we describe our past work using these methods, and make clear how they facilitate fruitful work in experimental philosophy. Section 4 contains a novel experiment on the emergence of human communication, including background theoretical work and a detailed presentation of our experimental design and results. In section 5 we conclude by addressing the philosophical upshots of the experiments presented here, and discussing, more generally, how economic methods can be fruitfully used to explore areas of philosophy involving strategic interactions.

2. Experimental Economics

The methods we present here are derived from experimental economics, a discipline which dates back to the middle of the last century. (See, for example, Allais (1953).) As in the other social sciences, experimentation in economics has allowed scholars to investigate human behavior in a highly regulated environment that controls for confounding factors, and thus to test and update theoretical predictions in the social sciences. The large body of work that has emerged addresses a range of phenomena spanning almost all imaginable strategic behavior in humans: cooperation, bargaining, altruism, coordination, price setting, norms, communication, market behavior, etc. (Kagel and Roth (2016) give an overview.)

For an example, Guth et al. (1982), in a seminal paper, have subjects play what is called an ultimatum game. One subject is given a set amount of money (say $10) and told that they may offer some of this to a second subject. The second subject has the option to accept what they are offered, or to reject it, in which case neither subject gets any money. The surprising result, which has since been widely replicated, is that subjects do not behave according to what the best models of rationality would predict. These models suggest that smart subjects should offer as little as possible, and that their partners should accept these offers since otherwise they get nothing. Instead, subjects make substantial offers, and reject ones that are too small. This experiment thus supplemented economists' understanding of bargaining behavior, challenged theoretical assumptions about human rationality, and prompted further theoretical work explaining why humans would not behave rationally in an ultimatum game.

The most important cornerstone of this branch of experimentation is *induced valuation*, the practice of prompting subjects to make choices in strategic situations where their choices have real consequences that the subjects care about (Smith 1976). This is usually done by making the payments rendered to subjects dependent on their performance in a trial. In the Guth et al. (1982) experiment on ultimatum bargaining, for example, subjects actually received the amount of money they earned. If one subject offered $5, and the other subject accepted this offer, they both went home with $5. If a subject offered $2, and the other subject rejected it, they went home with nothing.

Why induced valuation? Economic theory says nothing about what subjects say they would do in some strategic scenario, it only makes predictions about what subjects would, in fact, do (Croson (2005)). To test such predictions, then, subjects must be induced to make real economic choices. A study that merely asked subjects what they would offer in an ultimatum game, and what they would accept would not test the behavior that economic theory is about. There are many areas of theoretical philosophy where, likewise, predictions address subject behavior, rather than self-reports about predicted behavior. While one might think these collapse, empirical evidence suggests that humans are often quite bad at accessing and reporting their own cognitive states and predicting their own actions (Wilson & Dunn (2004), Poon et al. (2014)). In such cases, philosophers would do well do focus on experiments using methods of induced valuation.

Another key aspect of this methodology is that experiments tend to be largely context free. Experimenters often present subjects with just enough structure and information to capture the strategic situation they wish to induce. The goal is to abstract away from framing and structural features that might systematically bias the behavior of subjects. Consider again the ultimatum bargaining study. One way to present this study is to simply tell the subjects what the strategic structure of the game is. Another way might be to concoct a story about partners with two skill sets necessary to complete a project, where one offers some portion of the proceeds to the other. This latter framing, though, might bias participants to give more since their cultural norms about fairness in the workplace now become relevant in determining their behavior. Likewise, in the studies we will present here, we avoid describing the study as about language or communication. Since all participants are language users, this could influence their behavior by, for example, prompting them to behave in helpful communicative ways, rather than in their own best interests (Bruner et al. (2018)). This is not to say that framing effects are unimportant, and, indeed, experimenters regularly add framing to their studies to see how this influences subject behavior. (For example, Leliveld et al. (2008) find that subjects act differently in the ultimatum game depending on which subject is described as 'owning' the money. See also Fagley & Miller (1997).) The point is that these additions should be deliberate so that experimenters gain control as to how various additions to their paradigm impact subjects. Again, this practice is relevant to experimental philosophy, which often depends on highly specific vignettes or cases to test the intuitions of subjects.

One last standard practice in experimental economics is to avoid deceiving subjects. This is so that experimenters maintain control over subject expectations and motivations. If subjects have previously been deceived, or know that deception is possible in such experiments, they may not trust the experimental set-up they are presented with. For example, subjects may believe experimenters will secretly rig outcomes so as to pay out the least possible amount (Cooper, 2014). In the ultimatum bargaining game, subjects making offers might believe, for example, that there is no second subject, and experimenters will always pretend this second subject rejected the offer. The data gathered from such subjects will fail to track what experimenters are trying to test. For experimental philosophers who adopt the practice of induced valuation, adhering to the no-deception rule, and making sure subjects are aware of this, will help ensure that subjects are motivated by payoffs in the right ways.

The detailed examples we will present here are specifically within the realm of game theoretic and evolutionary game theoretic experimentation. Game theory is the study of *games* – simplified models of strategic interactions between humans. A game is specified by four elements: who interacts (*players)*, what they may do (*strategies*), what players get for various combinations of strategies they might play (*payoffs*), and what players know about the game (*information)*. In the ultimatum game, the two players are the subject who makes the offer and the one who decides whether to accept or reject it. The strategies are different for the two players - possible monetary offers for the first, and the choice to either accept or reject given various offers for the second. The payoffs are the amounts of money they get for some combination of these strategies - $5 each for an offer of $5 and an acceptance, for example. And the information in this case is that they know who the players are, what the strategies are, and what the payoffs are, i.e., they have full knowledge of the game.

Classic game theory uses these models, plus assumptions about human rationality, to predict and explain strategic behavior. These assumptions are, basically, that each player will try to maximize the amount of money they take home given what the other players are doing. Experiments are often useful in showing where such predictions do or do not hold, as with the ultimatum game.[1]

Evolutionary game theory, on the other hand, as applied to human culture, is the study of how humans learn and culturally evolve to deal with strategic scenarios as modeled by games. These models typically take a group of actors playing a game and add *dynamics* - rules for how their strategic behavior will change over time as a result of learning and cultural evolution. For example, such a model might involve a group of humans all playing the

[1] For another example, work on the famous prisoner's dilemma game has consistently shown that humans have a preference for altruistic behavior that does not accord with the predictions of rational choice (Sally 1995).

ultimatum game again and again. The dynamics might assume that each individual has a preferred strategy (maybe offer $4, and accept any offer over $3), but that each time they play there is some chance that they switch strategies and copy the most successful individual in their group. The model then sees what happens over time if everybody in the group learns via this rule. As such, evolutionary game theory makes predictions and provides explanations about how groups of humans will come to behave in strategic learning scenarios.

Predictions from these cultural evolutionary models are often different from those derived in classic game theory. For example, with respect to the ultimatum game, Skyrms (2014) shows how in an evolutionary model, people will sometimes learn to make high offers and reject ones that are too low, providing an explanation for laboratory behavior that previous models did not. Such cultural evolutionary predictions can themselves be tested by having groups of individuals play a game repeatedly in the lab and seeing what behavior emerges. In sections 3 and 4 we will give several examples.

Wherever philosophy uses game theory and evolutionary game theory, and wherever it makes predictions, or offers explanations of, strategic human interactions including communication, coordination, altruism, cooperation, social dilemmas, social norms, and resource distribution, the experimental methods we outline here can be of use. To date, there is a small literature in philosophy demonstrating just this point. In political philosophy experimental methods from economics have tested claims in the social contract tradition (Powell and Wilson 2008 and Smith, Sharbek and Wilson 2012, Bruner 2018). These methods are particularly apt as they allow one to explore behavior in the many hypothetical scenarios contract theorists have utilized to justify a variety of social arrangements[2]. In social philosophy, Devetag, Hosni and Sillari (2013) as well as Guala (2013) have used economic methods to probe issues relating to conventions and common knowledge, while Bicchieri and Lev-On (2007), Bicchieri and Xiao (2009), and Bicchieri and Chavez (2013) have used economic experiments to help develop and defend Bicchieri's influential account of social norms. Within epistemology Koppl et al. (2008) and Jonsson et al. (2015) have designed experiments to explore the ways in which group structure makes for better or worse epistemic groups.

In addition, techniques from experimental economics have been used to reinforce previous findings from more traditional experimental philosophy. Utikal and Fischbacher (2014), for instance, identify a version of the side-effect effect in an economics-style experiment. And Gold, Pulford and Colman (2014) conducted a 'real-life' version of the trolley problem (involving financial losses) and found reactions in this economics experiment were similar to reactions to the more hypothetical cases common in the philosophical literature.

3. Previous Results and Theoretical Grounding: the evolution of meaning

---

[2] For more on the relationship between ethics and experimental economics, see Ernst (2007) and Güth and Kliemt (2017). Also, see Guala (2005) for a discussion on what philosophers of science can learn from experimental economics.

Recently, philosophers and social scientists have developed a huge empirical, experimental and theoretical literature on the evolution of communication and language.[3] This social-scientific exploration dovetails with more traditional philosophical work regarding the meaning of linguistic terms, as well as debates over the conventionality of meaning. Quine, for instance, famously argued that conventionalist accounts of meaning are circular, as conventions themselves appear to presuppose language and meaning. In response, David Lewis developed a game-theoretic account of convention that was then used to demonstrate how, sans explicit agreement, linguistic terms can acquire meaning (Lewis, 1969). In particular, Lewis develops a novel communicative game, a *signaling game,* and argues that messages in this strategic scenario can acquire conventional meaning.

Lewis's signaling game has proved an extraordinarily useful framework for explorations of language and communication in philosophy. Brian Skyrms, for instance, has used signaling games to develop a notion of 'informational content,' a generalization of the more familiar propositional content (Skyrms, 2010a; Birch, 2014). Signaling games have also been used by philosophers to work out accounts of deception appropriate for non-intentional organisms (Fallis and Lewis, 2017; Martinez, 2015). Moreover, signaling games have been used to frame a variety of issues in the philosophy of biology relating to the evolution of language and proto-

[3] Biologists, too, have for some time pondered the evolution of language. John Maynard Smith and Eors Szathmary, for instance, went as far as to view the evolution of language as one of the major evolutionary transitions (Maynard Smith and Szathmary 1995).

language (Sterelny, 2012a,b). Outside of the philosophy of language, signaling games have been used in social epistemology because they allow philosophers to better understand the conditions that allow for the transfer of information among peers (Godfrey-Smith and Martinez, 201; Skyrms 2010b).

We now describe the communicative game introduced by Lewis. Lewis considers a simple strategic setting involving two players, a *sender* and a *receiver*. The sender is able to observe what *state* the world is in. These states might be, for example, that it is either sunny or raining. The sender can then select a *signal (*or *message)* from some available set to relay to the receiver. This set might be the words 'sunny' and 'raining', or the sounds 'bleh' and 'schmorg', or lighting either one or two lanterns in a belfry. Upon receipt of this signal, the receiver then picks an act to perform. It is assumed that certain acts performed by the receiver match particular states of the world and that both sender and receiver prefer the receiver perform the act that best matches the underlying state of the world. In this example, the acts might be to get out either umbrellas or sunscreen. Since both actors would like on the receiver to grab sunscreen if it is sunny, and umbrellas if it is raining, their interests in this example completely align.

Lewis considered the simplest possible version of this game – one with two possible states, two possible signals (or messages), and two possible actions. His key observation was that such models have two *signaling systems*. These are strategies where the sender always sends one signal in state 1 ('bleh' when it is raining, for example) and a different signal in state 2 ('schmorg' when it is sunny), and the receiver uses this regularity to perfectly coordinate action with the world (bring umbrellas in the rain and sunscreen in the sun). These systems are conventional, since either will do equally well as a communication system. In other words, it does not matter to the actors whether they use 'bleh' to mean rain or sun, as long as they coordinate. These systems are also stable in the sense that once actors have settled on one they will have no incentive to change their behavior. In this way, a signaling system is what game theorists call an *equilibrium* – a social arrangement where none of the parties can switch behavior and get a better payoff. Moreover, this particular equilibrium is optimal in the sense that it allows the players to get the highest payoffs possible, by always coordinating their behaviors with the state of the world. Figure 1 shows one of these signaling systems – where upon observing state 1 the sender sends message 1, which induces act 1. The other signaling system would match message 1 to state 2 and message 2 to state 1.

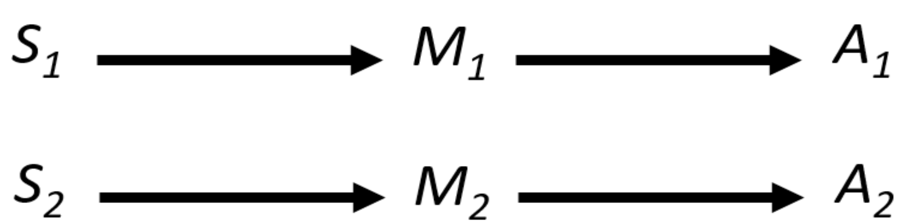


Figure 1: One of two signaling systems in David Lewis's signaling game with two states (S1, S2), two messages (M1, M2) and two acts (A1, A2).

Brian Skyrms (1996, 2010) was one of the first to explore these common-interest signaling games in an evolutionary context. This program was in part motivated by the fact that while Lewis' game-theoretic approach could explain the persistence of linguistic conventions, the account developed by Lewis did not provide a satisfactory explanation of the origin of these conventions. Using evolutionary game theory, on the other hand, Huttegger (2007) and Pawlowitsch (2008) show that in Lewis's version of the game, assuming the states are equally likely to occur, signaling systems are guaranteed to emerge under reasonable assumptions about how actors learn or evolve. In the words of Skyrms (1996), in these simple evolutionary models, "The emergence of meaning is a moral certainty" (93).

If the underlying signaling interaction is more complex – including, for example, more states of the world – it is possible for an evolutionary process to result in suboptimal outcomes where the sender sends the same signal for multiple states of the world. For instance, if the simple signaling game considered above is modified so that state 1 occurs most of the time (i.e., it is almost always rainy) the following arrangement is now stable: the sender sends one signal regardless of the state of the world and the receiver always performs the act appropriate for the more likely state (always brings umbrellas). This is an instance of what is called a *pooling equilibrium*. The sender's behavior is the same across both states of the world, and as a result the receiver is unable to glean information regarding the underlying state of the world by attending to the signal. As a result, the receiver essentially ignores the behavior of her counterpart and opts to take the act which is more likely to match the state of the world (see figure 2).

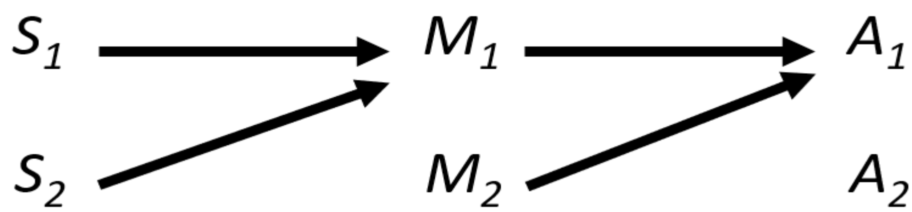


Figure 2: An example of a pooling equilibrium in a David Lewis signaling game.

When there are more states of the world, actors can also learn or evolve to send the same signal in several states of the world (but not all of them). These *partial pooling equilibria* emerge despite the fact that they, like pooling equilibria, are inefficient in that actors can do better by learning signaling systems. It may seem strange that these inefficient outcomes can culturally evolve. They are possible because despite their suboptimality given what the other player is doing, no one can do any better by trying another strategy.

Another complication emerges when the interests of sender and receiver diverge, such that they do not always prefer the same receiver action be performed in a given state. Imagine, for example, that the sender is a job candidate, and the state is either that they are capable or lazy. Suppose the receiver is an employer trying to glean information about the candidate. If the candidate is capable, the employer would like to hire them, otherwise not. The candidate, on the other hand, would always like to be hired. In other words, their interests line up only if the state is that the candidate is capable.

In such cases, signaling systems can often emerge, but only when messages are costly to send and only if the cost of a message depends in part on the underlying state of the world or the type of sender (Spence, 1973).[4] In the example just given, the costs of earning a college degree might be much lower for a capable candidate than a lazy one, so if the employer receives a 'signal' of a college transcript, they can deduce that the sender will be a capable employee.

Wagner (2013) showed that if this incentive structure is slightly modified to allow for slightly less costly messages, then a partially informative signaling system is possible (often referred to as the *hybrid equilibrium*).[5] In this case, capable candidates might always go to college, but lazy ones sometimes too, because the costs are a bit lower. As a result, the employer is not able to perfectly identify the quality of the candidate upon receiving a signal, although they do considerably better than chance. As a result, they sometimes, but not always, hire the candidate upon seeing a college transcript. Together, these insights form the basis of what is often referred to as *costly signaling theory*, which has been employed throughout the

---

[4] This is a bit of a simplification, since there are other ways to ensure communication in partial-conflict of interest settings. For alternative ways of ensuring communication, see Crawford and Sobel (1982), Akerlof (1970), Viscusi (1978), Grossman (1981), Milgrom (1981) and Jovanovic (1982).

[5] See also Huttegger and Zollman (2010) as well as Zollman (2013). For other experiments investigating conflict of interest signaling games (but not the hybrid equilibrium) see Cai and Wang (2006), Dickhaut, McCabe and Mukherji (1995), Forsythe, Lundholm and Rietz (1999) and Blume, Dejong, Kim and Sprinkle (2001).

social and biological sciences in order to explain a variety of initially puzzling signaling behaviors such as signaling by potential mates in the biological world, people on first dates, teenagers and their parents, students and their teachers, etc.

Much attention has been devoted to better understanding when and under what circumstances signaling systems will be likely to emerge in these games.[6] In sections 3.1 and 3.2 we describe laboratory experiments designed to test predictions which originate from this rich theoretical literature. See Blume et al. (2017) for a survey of related experimental literature.

3.1 David Lewis in the Lab

How likely is it that actual subjects will learn to assign meaning to initially meaningless symbols, successfully using these symbols to communicate with each other? And how does making the task more complex affect the likelihood of the emergence of meaning? Bruner et al. (2018) report the results of a laboratory experiment designed to explore these questions when the interests of sender and receiver coincide. Each run of the experiment proceeded as follows.

[6] See, for instance, Huttegger et al. (2010), Wagner (2009), Barrett and Zollman (2009), Skyrms (2012), and Brusse and Bruner (2017) for common interest signaling games and Wagner (2011, 2013, 2014), Zollman, Bergstrom and Huttegger (2013), Huttegger and Zollman (2010), Bruner, Brusse and Kalkman (2017), Bruner and Rubin (forthcoming), Bruner (2015), Huttegger, Bruner and Zollman (2015), Kane and Zollman (2016) and Martinez and Godfrey-Smith (2016) for work on conflict of interest signaling games.

A total of 12 subjects were recruited to the Experimental Social Science Lab at UC Irvine where they interacted anonymously via individual computer terminals. Six of these subjects were randomly assigned to be ‘senders,’ while the remaining six were ‘receivers.’ In order to avoid context effects, along the lines of the experimental methods described in section 2, the labels ‘sender’ and ‘receiver’ were replaced by the neutral labels 'role 1' and 'role 2'.

The experiment consisted in sixty rounds. During each round, each sender was randomly matched with a receiver. The sender was then randomly shown one of two symbols (# and * for example), intended to represent the state of the world. Upon observing the state symbol, senders then selected one of two different signal symbols (@ and ^ for example) to relay to the receiver. (These random symbols were intended to prevent actors from using salience clues to choose which signals matched each state.[7] In addition, notice again the completely context-free set up.) The receiver, upon observing only the signal, would then guess which state symbol the sender saw. At the end of each round both sender and receiver were told what symbol was initially presented to the sender as well as the receiver’s guess. Subjects received $1 USD for each out of four randomly chosen rounds where receivers guessed correctly, as well as a show-up fee of $7. (This randomization helps prevent wealth effects from influencing later rounds of experimentation (David and Holt, 1993).) Subjects were made aware of the payment structure and the structure of the signaling game they played at the

[7] Mehta et al. (1994) find that saliency can impact coordination behavior in game theoretic experiments, so we made every effort to reduce the saliency of any signal for any state.

beginning of the experiment. Since we were testing evolutionary predictions, we did break from standard economic practice by providing subjects with less information about population structure and play of their peers than is typical.[8]

Notice that this set-up embodied induced valuation in that actors were incentivized to signal in hopes of earning payoffs for coordination. It used minimal framing; presenting the signaling game without even using the language of signaling. And it avoided deception by making subjects aware of the strategic scenario they would face, and their potential payoffs.

In the game involving just two states of the world, two signals and two possible acts Bruner et al. (2018) find that, consistent with theoretical predictions, small groups tend to learn signaling systems when the states are equally likely. Subjects converged on a signaling system rather rapidly, usually taking less than 20 or 30 rounds (out of a total of 60). We also find that pooling behavior becomes increasingly likely as one state becomes more probable than the other. This, again, is consistent with theoretical predictions. We also considered a signaling game involving three states, three signals and three possible receiver responses. In the laboratory setting subjects often developed behavior that mimicked the expected partial-pooling outcomes, although observed play frequently resulted in a signaling system as well. In

[8] Blume et al. (1998), for example, in a similar experiment , gave all subjects a running history of play of all subjects (not just their interactive partners) in each round, which influenced the way subjects could learn. In order to more closely fit day to day learning environments, in Bruner et al. (2018) we did not provide such information.

sum, the behaviors of lab subjects showed just how easy it is to develop common interest signaling in a lab group (extending results from Blume et al. (1998)), and also that evolutionary game theoretic predictions are, indeed, reflected in the behaviors of humans learning to signal in the lab.[9]

3.2 Communication without the Cooperative Principle

Rubin et al. (*manuscript*) use similar methods to investigate the emergence of communication when the interests of the sender and receiver are not perfectly aligned, i.e., in situations like the employer-job candidate interaction described earlier. In particular, receivers tried to guess the 'type' of the sender (high or low). As described, senders always want the receiver to guess they are a 'high' type, while receivers want to correctly guess what type the sender is. One standard question here is whether different costs for high and low types can facilitate the evolution of honest signaling. Remember that theoretical work predicts when these costs are small a hybrid equilibrium, with partially honest communication, will emerge. Rubin et al. (*manuscript*) test whether subjects learn this sort of partial communication transfer in the lab.

[9] Of course, learning to communicate with only 2 or 3 signals is just a simple starting place. To get anywhere close to explaining human communication, we need to also talk about things like learning new signals (see, e.g. Alexander, Skyrms, and Zabell (2012)) and syntax (see, e.g. Barrett (2009)).

In this experiment, senders could choose to either pay a cost to send a signal or pay nothing and not send the signal. Senders were divided into high and low types, where the high type paid less to signal. To keep the language neutral, these types were referred to as blue and red, respectively. There were two treatments: a control treatment, where, as a result of signaling costs, the hybrid equilibrium did not exist, and an experimental treatment where it did. (See Rubin et. al. (*manuscript*) for details of the payoff structure and more details of the experimental set-up.)

Rubin et al. (*manuscript*) found that the experimental results were largely consistent with theoretical predictions. This was done by first comparing the control treatment to the experimental treatment, to see whether there was less information transfer in the experimental treatment as expected. Then, we checked whether there was still some information transfer of the type expected in the hybrid equilibrium: in particular, that the signal increased the likelihood that the sender was a high type. Again, these results confirmed the success of evolutionary game theoretic predictions as applied to human communication.

In both of the studies just described, of course, subjects learn to communicate in situations that are quite removed from the everyday communication of humans. Subjects interact in small groups, over computers, with highly restrictive rules of interaction and artificial incentives. One might wonder whether such studies can, indeed, tell us much about the emergence of real language. Notice, first, that the restrictive lab environment allows for a kind of conclusion that is not possible in more ecologically valid explorations. In particular, we are able to much more accurately assess whether humans learn the kind of behavior that game theoretic and evolutionary game theoretic models predict, and thus validate these models. As far as the real world, these experiments can still tell us plenty despite their artificial features. In the first case, contra Quine, conventional meaning can emerge, with no meta-level discussion, very easily. In the second case, as we argue in Rubin (*manuscript*), human communication is often assumed to be fully cooperative, but we show that communication can still occur despite persistent conflicts of interest.

4. Sim-Max Games and the Evolution of Categories

The new experiment we present here looks at the emergence of communication in a variation of the Lewis signaling game called the sim-max (similarity maximizing) game. This model was introduced by Jaeger (2007), and has been used by him and others to study the evolution of categories, both linguistic (Jaeger (2011)) and cognitive (Jaeger (2007), O'Connor (2014b)), the evolution of vague terms (Franke et al. (2010), O'Connor (2014a), Franke & Correia (2016)), the emergence of linguistic ambiguity (O'Connor (2015a)), and natural kind terms (O'Connor (forthcoming)).[10]

The sim-max model adds structure to the basic signaling game by assuming similarity relationships between states of the world. In particular, states are arrayed in a space where distance in the space tracks similarity. For instance, figure 3 shows two possible state spaces

[10] This exploration connects to previous work in economics both on common interest communication and on vagueness and ambiguity. See Blume et al. (1998, 2001, 2017) and Lipman (2009).

for such a game, with five states on a line and 9 in a plane. In figure 3.a, state 1 is more similar to 2 than it is to 4, say, because they are closer in the state-space. This similarity is hashed out in the payoffs of the game. It is assumed that each state has some ideal action that will yield the highest payoff. Actions are also successful, though less so, in states similar to the ideal one. Suppose that the five states in 3.a represent five levels of rain from completely sunny to a downpour. In state 5, the downpour, the ideal action might be to wear galoshes and a raincoat. In state 4, the heavy rain, this would also be a perfectly fine action because the states are similar, even though it might be strictly better to bring an umbrella rather than a heavy coat. In particular, these games usually assume that payoff is strictly decreasing as actors take actions that are less appropriate to the state of the world.

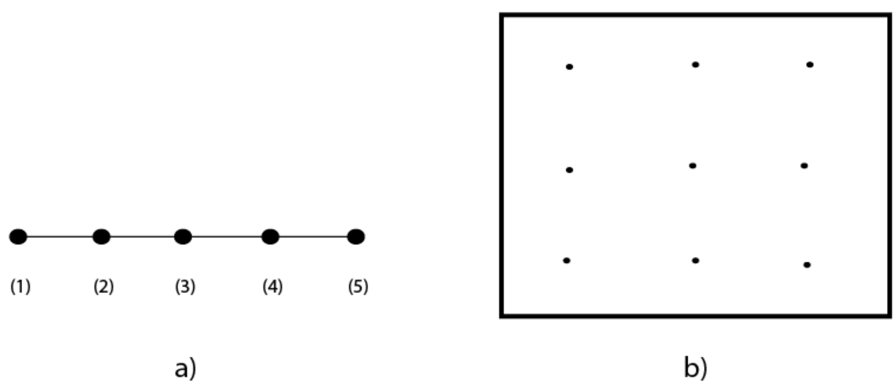


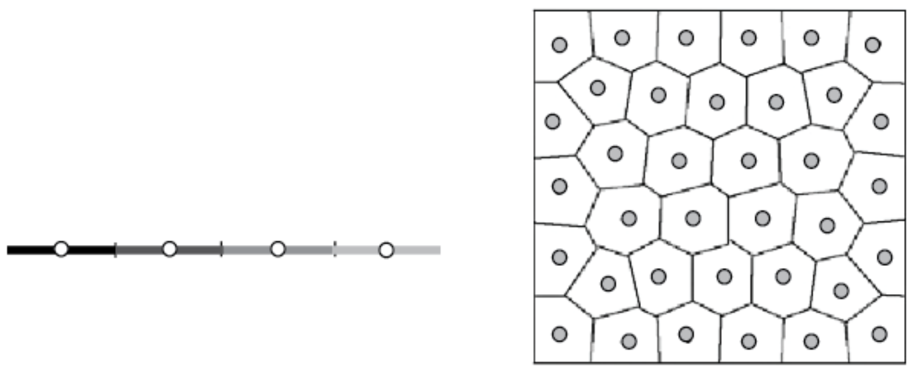
Figure 3: Two state spaces of sim-max games with a) five states arrayed on a line and b) nine states arrayed in a plane.

Play of the game is otherwise just like the Lewis signaling game. A sender observes the state of the world and sends a message. The receiver gets the message and chooses an action conditional on it. There is complete common interest between the actors, meaning that they always get the same payoff.[11] One typical assumption in these games is that the sender and receiver have access to fewer messages than there are states of the world. This means that they must use the same message for multiple states, i.e., develop communicative categories. If we take the states to be different levels of rain, and the actions to be appropriate responses, this would correspond to a situation where actors use just the terms 'rainy' and 'sunny' to communicate about a world that is in fact much more complex. This extends the Lewis model, where states are pre-specified, by inducing a strategic situation where agents must develop their own sets of states to apply terms to.

Previous results have shown that the categories we should expect to evolve in these games are (more or less) the optimal categories. (Jäger et al. (2011) call these categories *Voronoi languages*, after Voronoi tessellations in mathematics.) An optimal categorization will minimize, on average, the distance between the state of the world and the act taken, since this maximizes payoff to the actors. To do this, senders should use categories that are about equally sized, and receivers should respond with an action appropriate to a central, prototypical state in the category. Figure 4 shows examples of Voronoi languages for two state spaces, a) a line and b) a plane. Cells represent categories and open dots represent the action taken by the receiver in response to each category (not states as in figure 3). Returning to the rain example,

[11] If interests diverge, the game is more like the well-studied Crawford-Sobel signaling game in economics (Crawford & Sobel 1982).

an optimal strategy of this sort might be to use the term 'rainy' for any state from a downpour to a mild drizzle and the term 'sunny' for the rest of the states. Such a communicative strategy allows actors to achieve a high level of successful action, without having to develop a term for every possible state.



a) b)

Figure 4: Two voronoi languages. Cells represent categories and open dots the receiver response to a category. A) shows four categories in a linear state space, and b) 37 categories in a plane.

Jäger (2007) argues that these are the types of languages we should expect to evolve, though Elliott Wagner (personal correspondence) has shown that this may not always be the case and O'Connor (2014a, 2017) has found that under some learning rules categories that are similar to, but do not exactly correspond to Voronoi languages emerge. In particular, the categories may not exactly equally divide the state space, but nearly so. The slightly handwavy take-away is that in general we should expect actors in these games to develop categories that look more or less like the ones in figure 4, i.e., to use terms to represent sets of similar states.[12]

In addition, previous work has shown that actors in sim-max games have an advantage when it comes to learning to signal in that they can generalize lessons they learn over multiple states. Consider, for example, a standard signaling game (with no similarity structure) with 100 states. Actors must develop conventions for how to signal in every state anew, which is difficult and time consuming. In sim-max games, on the other hand, actors who learn a lesson in state 5 (bring raincoats) can apply that lesson to similar states that they have never encountered, speeding learning and improving payoffs (O'Connor (2014a), O'Connor (2015b)). And in addition, they may be able to avoid sub-optimal equilibria through this sort of generalization (Franke & Correia (2017)). This may help explain how real actors manage to successfully signal about so many real-world states: the structure of the world helps, by providing natural similarity classes over which to generalize learned linguistic conventions.

In the following we use experimental work to ask: do actors playing real sim-max games develop convention categories that facilitate information transfer? Are these categories optimal or near-optimal? And, do they generalize learning so as to improve their communicative success and speed of learning?

---

[12] O'Connor (2015a) gives a much more detailed overview of evolutionary predictions in these games and the work of Jäger (2007) and Jäger et al. (2011).

4.1 Experimental Design

The subjects consisted of undergraduate and graduate students from the University of California, Irvine recruited from the Experimental Social Science Laboratory subject pool. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007). As in Bruner et al. (2018) and Rubin et al. (*manuscript*), subjects interacted over 60 rounds. However, while subjects in Bruner et al. (2018) and Rubin et al. (*manuscript*) were matched randomly within a group of 12 every round, subjects in this experiment interacted with the same partner throughout. This made it easier for subjects to learn signaling behavior quickly, allowing us to use data from earlier rounds of the experiment.[13] This early data was crucial since we looked at games with many states, and thus needed a large number of data points to detect signaling patterns over these states.

At the start of each session, experimental subjects were asked to sit at a randomly assigned computer terminal. As in Bruner et al. (2018) and Rubin et al. (*manuscript*), subjects were given information about the strategic situation in a manner that was as context-free as possible. After every round, subjects were shown the state of the world, the signal sent, the receiver's action, and their own payoff for that round of the experiment. Each run of the

[13] We use data from rounds 20-60, compared with Bruner et al. (2018) and Rubin et al. (*manuscript*) who used data from rounds 50-60 to test convergence to equilibria.

experiment consisted of two treatments (so as to gather more data points given limits on time and resources).[14]

	2 Signals	3 Signals
Unstructured	2x2	3x3
	100x2	100x3
Structured	100x2 structured numbers	100x3 structured numbers
	100x2 structured colors	100x3 structured colors

Table 1: Summary of the different treatments.

There were 8 different treatments, summarized in table 1.[15] For each treatment, the first number is the number of states, and the second is the number of available signals. For the

[14] Because we had to reuse signals between the different treatments, we paired treatments that did not have an overlap in any of the signals to prevent using a signal in one treatment that had already gained meaning in a previous treatment. In each particular run, participants were given the same treatments in the same order. The order of these pairs were reversed across runs, e.g. one run had the 2x2 treatment followed by the 100x3 structured numbers treatment while another had the 100x3 structured numbers treatment followed by the 2x2 treatment.

[15] In the 2x2 treatment there were 8 subjects (meaning there were 4 interacting pairs), the 3x3 had 10 subjects, the 100x2 had 12, the 100x3 had 10, the 100x2 structured numbers had 16,

games we consider, the number of actions is the same as the number of states. First, in order to investigate how structured state spaces might influence signaling behavior, we tested some standard Lewis signaling games, i.e., without structure to the state space, for comparison. The 2x2 and 3x3 (read 'two-by-two' and 'three-by-three') treatments involved the same Lewis-style signaling games explored by Bruner et al. (2018), except that subjects were now interacting in pairs rather than groups, and successful coordination was rewarded with 100 points (to be translated to money at the end of the trial). In the 100x2 and 100x3 treatments, senders were shown a number from 1 to 100 (representing the state) and had only 2 or 3 signals available to communicate the state of the world to the receiver. Upon receipt of a signal, receivers had to guess the state of the world by typing in a number from 1 to 100. If the receiver guessed the correct state of the world, both subjects received 100 points, otherwise they received 0 points. While this game sets subjects up for failure, it provides an important comparison to similar sim-max games, as we will see.

The rest of the treatments involved subjects playing sim-max games. The 100x2 and 100x3 structured numbers treatments were the same as the 100x2 and 100x3 treatments - senders encounter states 1-100, and have either 2 or 3 signals available to coordinate action. But the payoffs were such that close guesses still paid off. In particular, subjects lost 2 points for each number away from the actual state their guess was. For example, if the actual state

---

the 100x3 structured numbers had 30, the 100x2 structured colors had 22, and the 100x3 structured colors had 16.

was 20 and the receiver guessed 33, both sender and receiver would receive 74 points that round (100 - |20-33| x 2 = 74).

The 100x2 and 100x3 structured colors treatments were also sim-max games, formally identical to the numbers treatments, but where the state stimuli were colors instead of numbers. We displayed a line that faded from very light to very dark green and subjects were told that it was divided into 100 evenly sized parts. The senders were then randomly presented each round with a state, in the form of this color line with an arrow pointing to one spot, as shown in figure 5. They then chose one of their available signals. Upon receipt of the signal, the receiver was presented with the same color line, and clicked on state they thought had occurred. Again, they received 100 points for guessing exactly right and lost 2 points for each unit away from the real state. The goal with these treatments was to test whether a different presentation of the state space would influence communicative behavior.

Figure 5: Sample of the state chosen in the 100x3 structured colors treatment.

As in Bruner et al. (2018) and Rubin et al. (*manuscript*), the signals available to senders were meaningless symbols, chosen so as to minimize the chance of subjects importing any pre-established meaning (e.g. we did not use the '>' sign as a possible signal for the structured numbers treatments, as subjects might already associate this with larger numbers).[16] For the 2x2 and 3x3 treatments, we also chose meaningless symbols to represent the states of the world.[17] These symbols were presented in a random order each round to as to prevent ordering from allowing subjects to coordinate.

Subjects received a show-up fee of $7 for participating in the experiment. In addition, subjects were paid for 10 rounds of the experiment: 5 rounds were randomly selected from each treatment for payment, excluding the first 10 rounds of each treatment to allow time for learning. Each subject's score, in terms of experimental points, for these 10 rounds was summed, and subjects were paid $1 for every 100 points they earned (rounded up to the nearest dollar). Subjects were made aware of this payment scheme, and which rounds could be chosen for payment, in the instructions.[18]

---

[16] The available signals were as follows: ^ and + for the 2x2 treatment; %, * and # for the 3x3 treatment; " and \ for the 100x2 treatment; ", \ and : for the 100x3 treatment; ` and ~ for the 100x2 structured treatments; and ?, / and [] for the 100x3 structured treatments..

[17] These were as follows: $ and @ for the 2x2 treatment and !, @ and > for the 3x3 treatment.

[18] Subjects participating in the 100x2 and 100x3 treatments received a 'bonus' payment of $3. It is not a common practice in experimental economics to hand out bonus payments that are independent of subjects' performance, but this ensured that subjects participating in these treatments were paid fairly for their time compared with other subjects. They were not told

4.2 Results

In what follows, we collapse data for the structured colors and numbers treatments and talk just about 'structured' treatments, where subjects played sim-max games.[19] First, we compare the structured treatments to the unstructured treatments, to see whether adding structure to the state space improves subject learning. Then, we look within the structured treatments in order to see whether the subjects could be said to use categories, and how close they were to equilibrium predictions in sim-max models.

4.2.1 Comparison to unstructured treatments

Does the fact that the world has underlying structure make it easier for people to communicate? Below, we test whether adding structure to the state space can improve

---

about this bonus payment until after the experiment was completed, so it will not have affected their behavior in the experiment.

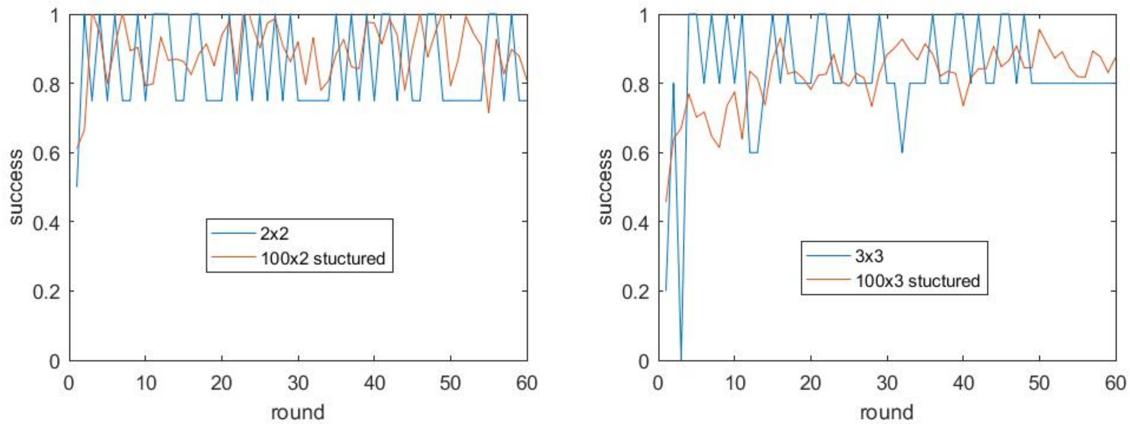
[19] These treatments were collapsed because the experiment was designed to test the effect of adding structure to the state-space. Behavior for the structured numbers and structured colors treatments was qualitatively similar and there were few significant differences between the two types of treatments. In the 3 signal treatments, subject's success rate was significantly higher for the structured colors (p=.0498), likely because there was significantly less variation in receivers' guesses (p=.02, see section 4.2.2 for a discussion). We will note places where this might influence our analysis, and show that it does not affect the conclusions we draw.

subjects' communicative success. Based on O'Connor (2014), we expect that success of subjects in the structured treatments will not be significantly different from the 2x2 or 3x3 treatments, but will be significantly different from 100x2 and 100x3 treatments. The idea is that receiving payoffs from being approximately correct can help subjects to reinforce categorization strategies and learn optimal signaling behavior. We follow O'Connor (2014) in using the following measure of success to compare how well learners are signaling across games where the base success rate is different:

$$\text{Success rate} = \frac{\text{average payoff in experiment}}{\text{expected payoff at equilibrium}}$$

Average payoff in the experiment was calculated using the points subjects earned in each round of the experiment.

Comparing the 2x2 with the 100x2 structured treatments, and the 3x3 with the 100x3 structured treatments, we used this measure and asked whether there was significantly less success in the structured treatments than in the standard Lewis signaling games. A one-tailed t-test revealed no significant difference in success rate (p=0.32 and p=0.50, respectively). As shown in figure 6, subjects reached their highest level of success very quickly, and there was little qualitative difference between the compared treatments.



(a) (b)

Figure 6: Success rates over time for a) the 2x2 versus the 100x2 structured treatments, and b) the 3x3 versus the 100x3 structured treatments. Success rates are averaged over all subjects.

Comparing the success rates of the 100x2 and 100x3 treatments to the structured state-space treatments is less straightforward. This is because the success rates in the 100x2 and 100x3 treatments could vary wildly if a receiver managed to guess the correct state a few times. That is, if the signal meant 'states 1-33' someone could by chance manage to guess exactly right twice in 10 rounds, giving them a much higher than expected payoff. Again using a one-tailed t-test, we found that for the 2 signal treatments, there was significantly less success for the unstructured versus the structured treatments for rounds 20-60 (p = .03). This confirmed our prediction that adding structure would improve learning for the subjects. However, for the 3 signal treatments, if we look at rounds 20-60, there is no significant difference (p = .41), but if we look at rounds 30-60, then there is (p = .001).[20] This was likely the result of a few chance correct guesses that happened in rounds 20-29 of the unstructured treatments.

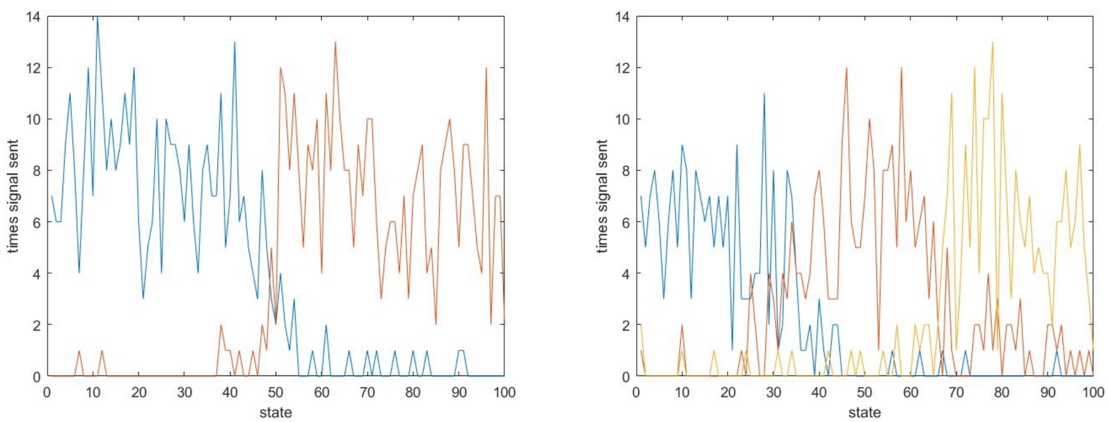
4.2.2 Categorization

When people have a limited number of signals to describe many possible states of the world, do they use categories in their communication? Do those categories divide up the world such that communication is as effective as possible? Jäger et al. (2011) and O'Connor (2014) predict that subjects will learn to use (approximate) Voronoi languages, where senders use categories that are about equally sized, and receivers respond with an action appropriate to a central, prototypical state in the category. We test this prediction in three parts below. First, we analyze sender behavior, then receiver behavior. Finally, we check that there was information transferred using the signals. For this analysis, we use data from rounds 20-60, because, as figure 6 shows, subjects had reached their maximum success rate by round 20 for both the 2 and 3 signal treatments. In other words, they had learned stable strategies by this point.

Are the categories approximately equally sized? Is each signal sent for approximately 1/n of the state-space (where n is the number of signals available)? This would involve, e.g., categorizing states 1-50 and 51-100 in the treatments with two signals, and categorizing 1-33, 34-66, and 67-100 in the treatments with three signals. As figure 7 shows, we observe a

[20] Since subject's success rate was significantly higher for the 100x3 structured colors (versus numbers) treatment, we also tested whether those had a significantly higher success rate than in the 100x3 treatment, and found that they did not (p=.36).

qualitative match with the prediction. To test this more rigorously, we used the following procedure: take the grouping implied by the senders' strategies (e.g., generally sending signal 1 for low states and signal 2 for high, etc.) to get an idea of what they take each signal to mean. Then recode so that signal 1 corresponds to the signal most frequently sent in the first 1/*n* of the state-space, etc. Then, assume they have divided the state-space into categories of size 1/*n* and look at proportion of time subjects sent the right signal in each category. Finally, check whether this is significantly less than 100% (as would be expected).



(a) (b)

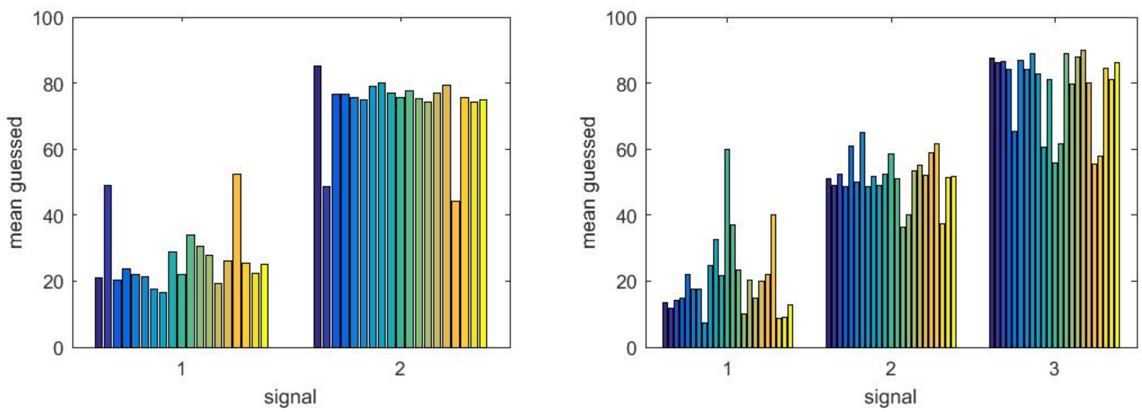
Figure 7: Senders' signaling behavior, averaged over all subjects, for a) the 2x2 versus the 100x2 structured treatments, and b) the 3x3 versus the 100x3 structured treatments.

Using a one-tailed t-test, we find that subjects' strategies are significantly different from equilibrium strategies of dividing the state space into categories of size 1/*n* (p = .015 when there are two signals and $p <.01$ when there are 3 signals). However, this is mostly due to subjects dividing the state space into categories of non-optimal size, rather than improperly signaling within the categories they have divided the state space into. In figure 7.a, for instance, if you look at when the signal meaning 'low' states is sent versus the one meaning 'high' states you can see that most 'mistakes' occur close to the boundary. This is mostly because different subjects drew the boundary between 'high' and 'low' at different places: for one subject high states might be from 45-100, for another 60-100. In fact, learning and evolutionary models often predict some conventionality as to where boundaries between categories are drawn, which accords with the behavior just described. This is, in part, because languages that are 'close' to Voronoi languages in that they draw the boundaries between categories near to the optimal spot are also very successful (O'Connor (forthcoming)). If we look at sender behavior away from the boundary between categories in the 2 signal case, the difference from 100% is no longer significant ($p = .067$), lending credence to the argument that conventionality of boundary position is causing deviation from expected behavior.[21]

Do the receivers take an action appropriate for a central, prototypical state in the category? That is, are the receiver's guesses in the middle of the 1/n sized categories? As a first check on whether this was the case, we measured the distance from the equilibrium strategy, assuming the sender uses categories of size $1/n$. The receivers' guesses are significantly off

[21] We did this by ignoring the middle states (from 36 to 64) and looking at behavior in roughly the top and bottom thirds of the state space.

from the equilibrium predictions (p << .001 in both cases).[22] After some initial learning, receivers' guesses are on average about 6 units (either in number or units of the color spectrum) away from the equilibrium prediction for both the 2 and 3 signal treatments. This is because there was high variance in receivers' guesses. So, for instance, if the equilibrium prediction was to guess 25, a receiver may have guessed 19 in one round, then 31 in the next, etc.



(a) (b)

Figure 8: Receivers' mean guesses for a) the 100x2 structured treatments and b) the 100x3 structured treatments. Each bar represents one subject's guesses.

---

[22] Though in the 3 signal structured colors treatment there was less variance in receiver guesses, these guesses were still significantly off from the equilibrium prediction (p<<.001).

We can see the various receiver strategies in figure 8, which shows what each receiver guessed after receiving each signal, averaged over rounds 20-40. As we can see, most subjects' strategies were on average close to the equilibrium prediction, but, as mentioned, their actual guesses tended to vary quite a lot.

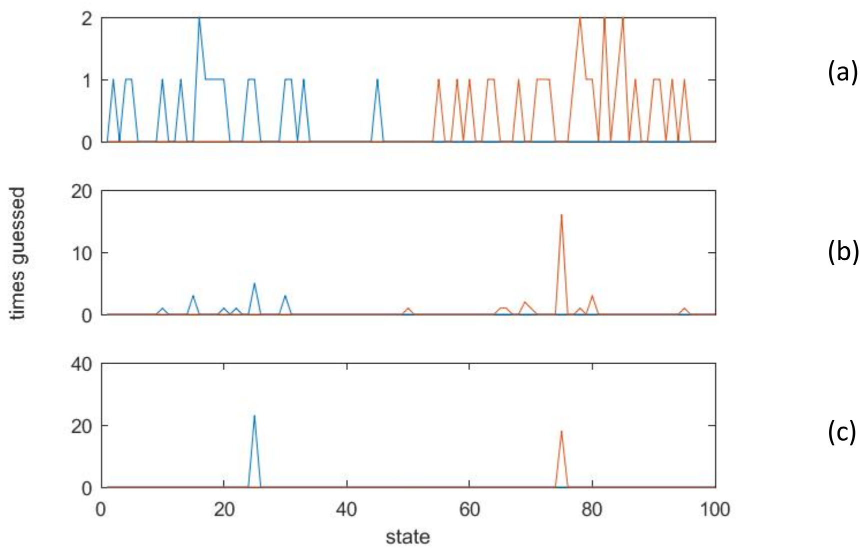
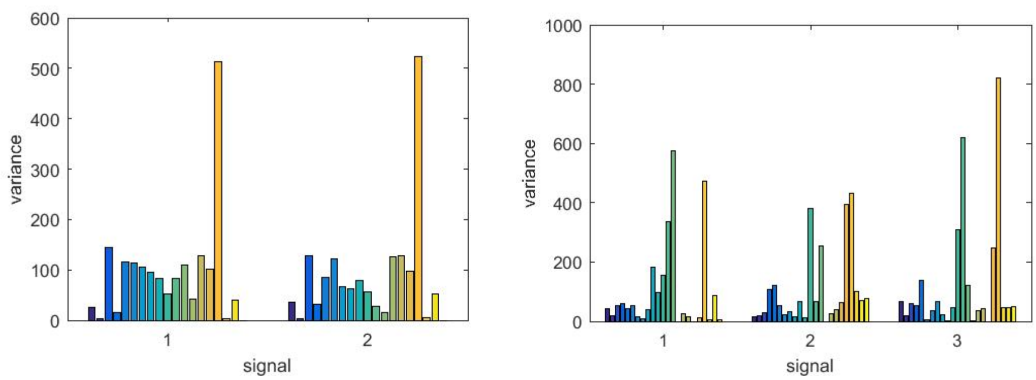


Figure 9: Sample receiver behavior from the 100x2 structured numbers treatment.

Figure 9 shows some of the receiver strategies from the 100x2 structured numbers treatment. Each of the three receivers guessed, on average, close to the equilibrium prediction, though only the receiver strategy in 4c represents an equilibrium strategy. More common were strategies like that shown in 4b, where receivers centered around the equilibrium strategy but often made guesses which were somewhat higher or lower. Receivers also occasionally employed a strategy where they divided the state-space into approximately 1/*n* sized categories, then guessed random states within each category, as shown in 4a. A summary of the variance in receiver strategies is shown in figure 10.



(a) (b)

Figure 10: Variance in receivers' guesses for a) the 100x2 structured treatments and b) the 100x3 structured treatments. Each bar represents the variance in one subject's guesses.

The variance in receiver guesses can help explain the ambiguous evidence found in section 4.2.1 for the fact that structure can aid in learning categories: even though subjects in the structured treatments learn to categorize, their success rate is lower than expected because they do not optimally respond to learning that the state of the world is in a certain category.

The variance in receiver guesses might be explained by a phenomenon similar to probability matching, which is a well-known phenomenon in experimental economics. When subjects employ a probability matching strategy, the frequency of their predictions of a state of the world matches the state's probability of occurring. For instance, if there are two states of the world and state one occurs 70% of the time, then 7 out of 10 times the subject is asked to predict the state of the world they will guess state one and the other 3 times they will guess state two. This happens despite the fact that the optimal strategy is to guess the more likely state every time. One explanation of this phenomenon is that subjects try to look for patterns, even when there are none, and predict the next state based on these patterns (Vulkan 2000). For instance, a subject may guess state two when they think it is 'due' to come up. Our subjects may have employed similar reasoning, making their guesses based on an anticipation of a particular state (within the range of states associated with a particular signal) fitting some pattern, rather than based on utility maximization, despite being told that states of the world were randomly determined by the computer.

Was there information transfer? Did the signals contain information about the state of the world? In order to test whether this was the case, we compared subject behavior to a null hypothesis of no information transfer. More specifically, we compared the average payoff of subjects in the experiment to the expected payoff subjects would receive if they were to ignore the signals, or if there were no information about the state of the world contained in the signal. If there was no information transfer, the best possible strategy would be to guess in the center of the state space (e.g. either 50 or 51 for the structured numbers treatments). This minimizes the expected difference between one's guess and the state of the world, and it yields an expected payoff of 50.[23] Using a one-tailed t-test, we found that subjects in the structured state space treatments did utilize the information content present in the signals to perform significantly better than if there were no information transfer ($p < 0.05$ in both the two signal and three signal cases).

5. Conclusion: Experimental Economics and Philosophy

The experiments described in this paper have implications for traditional questions in philosophy. Simple lab experiments completely invalidate theoretical arguments that symbolic meaning cannot emerge naturally (Quine (1936)) – on the contrary, it can emerge on the matter

of minutes. Furthermore, these experiments help validate the evolutionary models that philosophers have used to argue this point previously. Experiments on partial conflict of interest signaling show how groups of individuals can develop partially communicative conventions. This is relevant to claims in philosophy of language, especially in pragmatics, that human language is usually maximally helpful and informative (Grice (1991)). Game-theoretic models of sim-max games suggest the emergence of a categorization scheme that allows for informative communication. In line with the theoretical predictions, we show that near-optimal

[23] Subjects employing this strategy would be on average 25 units away, meaning their average payoff would be 100 - 25 x 2 = 50.

categories are in fact utilized by subjects. These findings have implications for understanding proto-typically vague terms, which look much like those that emerge in the experiments here. Despite the imprecision that has worried logicians and philosophers, these terms do a good job of helping humans communicate when the world is complex (O'Connor (2014), Franke and Correia (2016)).

Economics experiments of this kind have broad applicability across philosophy. Wherever philosophers examine the strategic interactions of people – be these cooperative, communicative, antagonistic, political, etc. – economics style experiments, and game theoretic experiments in particular can be illuminating. In particular, whenever experimental philosophers wish to learn about the actual behavior of subjects, rather than self-report, these methods can come into play.

Suggested Readings

Blume, A., Lai, E. K., & Lim, W. (2017). *Strategic information transmission: A survey of experiments and theoretical foundations*. Working paper. http://wooyoung.people.ust.hk/Info-Transmission-Survey-04-30-2017.pdf

Croson, R. (2005). The method of experimental economics. *International Negotiation*, *10*(1), 131-148.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics*. Princeton university press.

Friedman, D., & Sunder, S. (1994). *Experimental methods: A primer for economists*. Cambridge University Press.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, *66*(2), 274-279.

References

Akerlof, G. (1970). The market for "lemons": Quality uncertainty and the market mechanism. The Quarterly Journal of Economics, 84(3), 488-500.

Allais, P. M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats etaxiomes de l'ecole americane. *Econometrica*, *21*, 503–546.

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

Barrett, J. A. (2009). The evolution of coding in signaling games. *Theory and Decision*, *67*(2), 223-237.

Barrett, J. & Zollman, K. (2009). The role of forgetting in the evolution and learning of language. *Journal of Experimental and Theoretical Artificial Intelligence*, 21(4), 292-309.

Bicchieri, C., & Chavez, A. K. (2013). Norm manipulation, norm evasion: experimental evidence. *Economics & Philosophy*, *29*(2), 175-198.

Bicchieri, C., & Lev-On, A. (2007). Computer-mediated communication and cooperation in social dilemmas: an experimental analysis. *Politics, philosophy & economics*, *6*(2), 139-168.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, *22*(2), 191-208.

Blume, A., DeJong, D. V., Kim, Y. G., & Sprinkle, G. B. (1998). Experimental evidence on the evolution of meaning of messages in sender-receiver games. *The American Economic Review*, *88*(5), 1323-1340.

Blume, A., DeJong, D., Kim, Y., & Sprinkle, G. (2001). Evolution of communication with partial common interest. *Games and Economic Behavior*, 37, 79-120.

Blume, A., Lai, E. K., & Lim, W. (2017). *Strategic information transmission: A survey of experiments and theoretical foundations*. Working paper.

Bruner, J. (2018) Decision making behind the veil: an experimental approach. In Oxford Studies in Experimental Philosophy (eds. Tania Lombrozo, Shaun Nichols and Joshua Knobe), Oxford University Press.

Bruner, J., O'Connor, C., Rubin, H., & Huttegger, S. (2018). David Lewis in the Lab. *Synthese*, 195(2), 603-621

Bruner, J., Brusse, C., & Kalkman, D. (2017). Cost, expenditure and vulnerability. *Biology and Philosophy*, 32(3), 357-375.

Bruner, J., & Rubin, H. (forthcoming). Inclusive fitness and the problem of honest communication. *British Journal for the Philosophy of Science*.

Bruner, J. (2015). Disclosure and information transfer in signaling games. *Philosophy of Science*, 82(4), 649-666.

Brusse, C, & Bruner, J. (2017) Responsiveness and robustness in David Lewis signaling games. Philosophy of Science, 84(5), 1068-1079.

Cai, H., & Wang, J. (2006). Over-communication in strategic information transmission games. *Games and Economic Behavior*, 56, 7-36.

Cooper, D. J. (2014). A Note on Deception in Economic Experiments. *Journal of Wine Economics*, *9*(2), 111.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, 1431-1451.

Croson, R. (2005). The method of experimental economics. *International Negotiation*, *10*(1), 131-148.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics*. Princeton university press.

Devetag, G., Hosni, H., & Sillari, G. (2013). You better play 7: mutual versus common knowledge of advice in a weak-link experiment. Synthese, 190(8), 1351-1381.

Dickhaut, J., McCabe, K., & Mukherji, A. (1995). An experimental study of strategic information transmission. *Economic Theory*, 6, 389-403.

Ernst, Z. (2007). Philosophical issues arising from experimental economics. *Philosophy Compass*, 2(3), 497-507.

Fallis, D. and P. Lewis (2017). Toward a formal analysis of deceptive signaling. *Synthese*.

Fagley, N. S., & Miller, P. M. (1997). Framing effects and arenas of choice: Your money or your life?. *Organizational behavior and human decision processes*, *71*(3), 355-373.

Forsythe, R., Lundholm, R., & Rietz, T. (1999). Cheap talk, fraud and adverse selection in financial markets: some experimental evidence. *Review of Financial Studies*, 12, 481-518.

Franke, M., Jäger, G., & Van Rooij, R. (2010, November). Vagueness, signaling and bounded rationality. In *JSAI international symposium on artificial intelligence* (pp. 45-59). Springer, Berlin, Heidelberg.

Franke, M., & Correia, J. P. (2016). Vagueness and imprecise imitation in signalling games. *The British Journal for the Philosophy of Science*.

Friedman, D., & Sunder, S. (1994). *Experimental methods: A primer for economists*. Cambridge University Press.

Gold, N., Pulford, B., & Colman, A. (2015). Do as I say, don't do as I do: differences in moral judgments do not translate into differences in decisions in real-life trolley problems. Journal of Economic Psychology, 47, 50-61.

Gold, N., Pulford, B., & Colman, A. (2013). Your money or your life: comparing judgments in trolley problems involving economic and emotional harms, injury and death. Economics and Philosophy, 29(2), 213-233.

Grice, H. Paul. *Studies in the Way of Words*. Harvard University Press, 1989.

Grossman, S. (1981). The informational role of warranties and private disclosure about product quality. The Journal of Law and Economics, 24(3), 461-483.

Guala, F. (2005). The methodology of experimental economics. Cambridge University Press.

Guala, F. (2013). The normativity of Lewis conventions. *Synthese*, 190(15), 3107-3122.

Guth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*(4), 367–388.

Güth, W., & Kliemt, H. (2017). Experimental Economics—A Philosophical Perspective. *Oxford Handbooks Online.*

Huttegger, S. (2007). Evolution and the explanation of meaning. *Philosophy of Science*, 74(1), 1-27.

Huttegger, S., Skyrms, B., Smead, R., & Zollman, K. (2010). Evolutionary dyanmics of Lewis signaling games: signaling systems vs partial pooling. *Synthese*, 172(1), 177-191.

Huttegger, S., & Zollman, K. (2010). Dynamic stability and basins of attraction in the Sir Philip Sidney game. *Proceedings of the Royal Society B*, 94, 1-8.

Huttegger, S., Bruner, J., & Zollman, K. (2015). The handicap principle is an artifact. *Philosophy of Science*, 82(5), 997-1009.

Jäger, G. (2007). The evolution of convex categories. *Linguistics and Philosophy*, *30*(5), 551-564.

Jäger, G., Metzger, L. P., & Riedel, F. (2011). Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals. *Games and economic behavior*, *73*(2), 517-537.

Jonsson, M., Hahn, U., & Olsson, E. (2015). The kind of group you want to belong to: effects of group structure on group accuracy. Cognition, 142, 191-204.

Jovanovic, B. (1982). Truthful disclosure of information. Bell Journal of Economics, 13, 36-44.

Kane, P., & Zollman, K. (2016). An evolutionary comparison of the handicap principle and hybrid equilibrium theories of signaling. *PLoS ONE*, 10(9), e0137271. doi:10.1371/journal.pone.0137271.

Koppl, R., Kurzban, R., & Kobilinsky, L. (2008). Epistemics for forensics. Episteme, 5(2), 141-159.

Leliveld, M. C., van Dijk, E., & Van Beest, I. (2008). Initial ownership in bargaining: Introducing the giving, splitting, and taking ultimatum bargaining game. *Personality and Social Psychology Bulletin*, *34*(9), 1214-1225.

Lipman, Barton L. (2009) "Why is language vague?." *Boston University*. Unpublished manuscript.

Martinez, M. (2015). Deception in sender-receiver games. Erkenntnis, 80: 215-227.

Martinez, M., & Godfrey-Smith, P. (2016). Common interest and signaling games: A dynamic analysis. *Philosophy of Science*, 83(3), 371-392.

Mehta, J., Starmer, C., & Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, *84*(3), 658-673.

Milgrom, P. (1981). Good news and bad news: Representation theorems and applications. Bell Journal of Economics, 12, 380-391.

O'Connor, C. (2014). The evolution of vagueness. *Erkenntnis*, *79*(4), 707-727.

O'Connor, C. (2014). Evolving perceptual categories. *Philosophy of Science*, *81*(5), 840-851.

O'Connor, C. (2015). Ambiguity is kinda good sometimes. *Philosophy of Science*, *82*(1), 110-121.

O'Connor, C. (2015). Evolving to generalize: Trading precision for speed. *The British Journal for the Philosophy of Science*, *68*(2), 389-410.

O'Connor, C. (forthcoming). Games and Kinds. *The British Journal for the Philosophy of Science.*

Pawlowitsch, C. (2008). Why evolution does not always lead to optimal signaling systems. Games and Economic Behavior, 63, 203-226.

Poon, C. S., Koehler, D. J., & Buehler, R. (2014). On the psychology of self-prediction: Consideration of situational barriers to intended actions. *Judgment and Decision Making*, *9*(3), 207.

Powell, B., & Wilson, B. (2008). An experimental investigation of Hobbesian jungles. Journal of Economic Behavior & Organization, 66, 669-686.

Quine, W. V. (1936). Truth by convention.

Rubin, H., Bruner, J., O'Connor, C. & Huttegger, S. (manuscript) Communication without the cooperative principle.

Sally, D. (1995). Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Rationality and society*, *7*(1), 58-92.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge University Press.

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

Skyrms, B. (2010b). The flow of information in signaling games. Philosophical Studies, 147: 155. https://doi.org/10.1007/s11098-009-9452-0

Skyrms, B. (2012). Learning to signal with probe and adjust. *Episteme*, 9, 139-50.

Skyrms, B. (2014). *Evolution of the social contract*. Cambridge University Press.

Smith, A., Skarbek, D., & Wilson, B. (2012). Anarchy, groups and conflict: an experiment on the emergence of protective associations. Social Choice and Welfare, 38(2), 325-353.

Smith, J. M., & Szathmary, E. (1995). *The major transitions in evolution*. Oxford University Press.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, *66*(2), 274-279.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, 87(3), 355-374.

Sterelny, K. (2012a) *The Evolved Apprentice*. MIT Press.

Sterelny, K. (2012b) A glass half-full: Brian Skyrms' signals. Economics and Philosophy, 28, 73-86.

Utikal, V., & Fischbacher, U. (2014). Attribution of externalities: an economic approach to the Knobe effect. Economics and Philosophy, 30, 215-240.

Viscusi, W. (1978). A note on "lemons" markets with quality certification. Bell Journal of Economics, 9(1), 277-279.

Vulkan, Nir (2000). An economist's perspective on probability matching*. Journal of Economic Surveys, 14 (1)*, 101-118.

Wagner, E. (2009). Communication and structured correlation. *Erkenntnis*, 71, 377-393.

Wagner, E. (2011). Deterministic chaos and the evolution of meaning. *British Journal for the Philosophy of Science*, 63(3), 547-575.

Wagner, E. (2013). The dynamics of costly signaling. *Games*, 4(2), 163-181.

Wagner, E. (2014). Conventional semantic meaning in signaling games with conflicting interests. *British Journal for the Philosophy of Science*, 66(4), 751-773.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual review of psychology*, *55*.

Zollman, K. J. S., C. T. Bergstrom, and S. M. Huttegger (2013). Between Cheap and Costly Signals: The Evolution of Partially Honest Communication. *Proceedings of the Royal Society London, B*, 280, 20121878.