Abstract

People can instantaneously create novel conventions that link individual communicative signals to meanings, both in experiments and everyday communication. Yet a basic principle of natural communication is that the meaning of a signal typically contrasts with the meanings of alternative signals that were available but not chosen. That is, communicative conventions typically form a system, rather than consisting of isolated signal-meaning pairs. Accordingly, creating a novel convention linking a specific signal and meaning seems to require creating a system of conventions linking possible signals to possible meanings, of which the signal-meaning pair to be communicated is merely a sub-case. If so, people will not link signals and meanings in isolation; signal-meaning pairings will depend on alternative signals and meanings. We outline and address theoretical challenges concerning how instantaneous conventions can be formed, building on prior work on "virtual bargaining," in which people simulate the results of a process of negotiation concerning which convention, or system of conventions, to choose. Moreover, we demonstrate empirically that instantaneous systems of conventions can flexibly be created in a 'minimal' experimental paradigm. Experimental evidence from 158 people playing a novel signaling game shows that modifying both the set of signals, and the set of meanings, can indeed systematically modify the signal-meaning mappings that people may instantaneously construct. While consistent with the virtual bargaining account, accounting for these results may be challenging for some accounts of pragmatic inference.

Keywords: communication, conventions, joint inference, pragmatics, preemption, signalling games

Word count (main text): 17,600

Human communication can map signals onto meanings in astonishingly flexible ways, shaped by the communicative challenges of the moment (H. H. Clark, 1996; Grice, 1957; Levinson, 2000). Moreover, within a context, the mapping between possible signals and possible meanings is typically systematic. That is, signals are not interpreted in isolation, but by contrast with other signals that have been chosen instead. Thus, people appear to be able flexibly to create a systematic mapping from possible signals into possible meanings that is tailored to the current communicative context.

While driving, for example, we routinely send and receive signals such as left, right, or "null" turn-signal lights in situations like that shown in Figure 1a. At a particular moment, suppose that an automobile A can communicate its intentions with an on-coming cyclist C by indicating left (←), right (→) or not indicating (∅) (we ignore other modes of communication such as waves, headlight flashes or honks, and later possible signals). There are four possible turnings and a hazard blocking A's path (Fig. 1a).

How does C interpret these signals? Intuition from driving experience suggests something like the following. If all four turns are available, indicating left or right at this point suggests an intention to turn left or right before the hazard; no signal implies continuing past the hazard (Fig. 1b). But if one or both of the immediate turns is blocked (e.g., by a parked car), then the signals systematically change their meaning. For example, if the immediate left turn is blocked (Fig. 1c), then a left signal cannot imply an intention to turn down it; it must instead indicate an intention to move into the cyclist's path around the hazard—irrespective of later intentions. If the immediate right turn is blocked (Fig. 1d), then signaling right implies taking the

later turn (something specifically excluded before). If both paths are blocked but the main road is clear of hazards (Fig. 1e), indicating left now implies the intention to take the left turn up ahead. Notice, in particular, that the signals associated with the three routes *after* the hazard are systematically changed by the availability of turnings *before* the hazard. When we consider further subtleties, such as the possibility of signaling later, or using other signals such as waves, honks or headlights, the complexities multiply rapidly. Yet, for our driver and cyclist to avoid collision or standstill, they must reliably agree on these meanings in real time.

This example illustrates how changes in available options systematically change the entire mapping between 'signals' and 'meanings.' In previous work (Misyak, Noguchi & Chater, 2016), we have shown that the meaning attached to an individual signal can be modified 'in the moment' depending on communicative goals and partners' common ground. But examples such as the above illustrate that people systematically modify entire mappings of signals and meanings.

Such systematic remapping is widespread: indeed, a basic principle of communication is that the meaning of any particular signal typically contrasts with the meanings of alternative signals that were available, but were not chosen (E. V. Clark, 1988, 1990; E. V. Clark & H. H. Clark, 1979; Levinson, 2000). That is, communicative conventions form a system, rather than consisting of isolated signal-meaning pairs.

Many systems of communicative conventions are, of course, fairly stable across a linguistic community and across time, and are established through slow processes of cultural evolution (e.g., Brighton, Smith & Kirby, 2005; Christiansen & Chater, 2008). Moreover, a wide variety of experimental studies have shown that such conventions can be created rapidly, through

short periods of communicative interaction between two or more people (e.g., Clark & Wilkes-Gibbs, 1986; Garrod & Doherty, 1994; Hawkins, Frank & Goodman, 2020; Hawkins, Franke, Smith & Goodman, 2018).

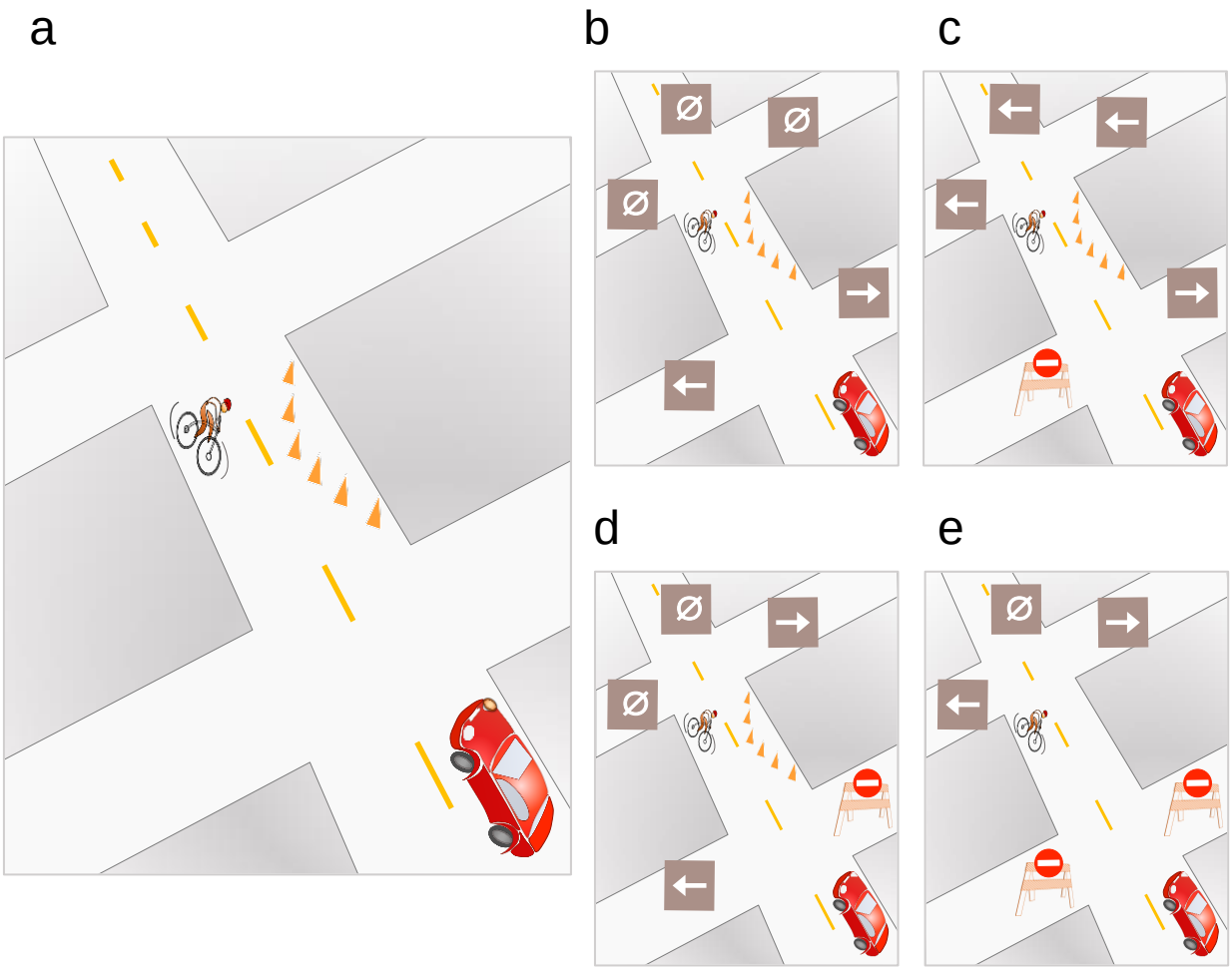

Fig. 1.

Systematic re-mappings of communicative signals in a driving example. An automobile (lower-right) communicates early its intended trajectory to an approaching cyclist through the use of a turn-signal (a). There are four possible turnings for the car's trajectory and an upcoming road hazard (marked by traffic pylons). The symbol in each box shows the turn-signal (← = left, → = right, ∅ = none) the car is likely to use if choosing the associated path. In the base case (b), all

possible turnings are accessible to the car. In other cases (c, d, e), barricades with "no-entry" signs represent inaccessible paths. As the set of available paths changes, so too does the interpretation of the car's signals—reorganizing as a system each time.

In this paper, by contrast, we focus on whether people can flexibly and instantaneously (i.e., in a single communicative interaction) modify an entire system of signal-meaning mappings in novel communicative contexts. Doing this successfully seems computationally challenging: the sender (here, the automobile A) and the receiver (cyclist, C) must attend to the range of available signals (here indicating left, right, or not at all) and meanings (here, concerning A's intended trajectory), and link them in a way that is optimized to the situation. Crucially, for communication to succeed, both sender and receiver must independently alight on the same system of conventions, and hence the same mapping between the specific signal and meaning in play in a communicative exchange. This means that, in order to flexibly impose a particular mapping between signals and meanings, both sender and receiver should be confident that the other will adopt the same mapping.

One possible claim is that such calculations are not made on-the-fly, but result from accumulated experience of similar driving situations. Being able to rely on prior experience relieves the burden of complex, on-line pragmatic computations. It may also reduce the problem of ensuring that both sender and receiver can be confident that they are using the same system of conventions, because they may automatically be aligned if they are drawing upon past experience from a common set of clear cultural norms. Indeed, in natural language, signal-meaning mappings are often modulated in standardized ways to maintain appropriate contrasts between distinct signals (e.g., Levinson, 2000). Thus, for example, saying *a few people like*

cinnamon is logically compatible with *everyone likes cinnamon*, but is typically used contrastively. In view of the impact of past conventionalization in natural language, it is not easy to determine how much inference is carried out in the moment.

In this paper, we explore the creation of instantaneous conventions, both theoretically and empirically. First, in *Instantaneous conventions, pragmatic reasoning and the virtual bargaining framework*, we clarify what it means for a convention to be instantaneous, and note that successful momentary communication requires that the sender and receiver must choose the same instantaneous convention out of many possibilities. This problem of choosing between multiple equilibria is not addressed by many theories of pragmatic reasoning, but can, we argue, be resolved by seeing aligning on a convention as arising from a process of "virtual bargaining," a process in which both sender and receiver simulate which convention they would choose, if they were able to negotiate. Second, in *Forming instantaneous systems of conventions*, we investigate how far people can flexibly modify systematic signal-meaning mappings in an experimental communication paradigm by which we explore whether people can alight upon instantaneous conventions in the simplest set-up we envisaged. People play a communication game, where both possible signals and meanings are very restricted; and in a domain which is abstract and unfamiliar, so that the influence of prior conventionalization outside the experiment is minimized. We test whether people are able instantaneously to make systematic shifts in signal-meaning mappings in a single communicative interaction.

1. Instantaneous conventions, pragmatic reasoning and the virtual bargaining framework

1.1. Instantaneous conventions

In his classic analysis of conventions, the philosopher David Lewis (1969) notes that people frequently face problems of coordination, which can be solved in many ways—what matters for coordination to be achieved is that people choose the *same* solution. Thus, driving on the left versus the right are equally good options a priori—but it is crucial that a given population of road users drives consistently on the same side of the road. Lewis argues that the function of conventions is to solve such problems. A convention is a self-reinforcing solution to a coordination problem: self-reinforcing in the sense that if others choose a particular convention (e.g., driving on the right) then each individual has an incentive to adopt that convention. The choice of convention may be somewhat, or even entirely, arbitrary; but by adopting a particular convention, road users are able to coordinate successfully, without continual deadlocks and collisions. In Lewis's analysis, conventions crucially solve repeated versions of the same coordination problem. Thus, each time two cars pass on the road, the same question arises concerning who drives on which side; and the convention solves this not just for the present but also future occasions.

How are we to apply this viewpoint to understand how people communicate so flexibly depending on the specifics of the situation that they face, such as using turn-signal lights in the example above (or indeed using natural language)? The standard viewpoint is to distinguish sharply between the *semantic* properties of a communicative system, captured by stable conventions across time and situations; and *pragmatic* reasoning which flexibly applies, modifies and perhaps goes beyond, these semantic conventions in the light of the specific communicative

context (e.g., Lascarides & Copestake, 1998).[1] From this point of view, the very idea of an instantaneous convention (Misyak, Noguchi & Chater, 2016) may appear misguided—conventions are seen as inherently stable across situations, and any situation-specific inference is viewed as non-conventional, pragmatic inference.

Yet this picture runs into substantial difficulties on closer inspection. For example, what is the conventional meaning of a left turn-signal light, independent of context? It is unclear that this is even a coherent question. We might attempt a gloss such as "I'm about to turn left," but, as the examples above indicate, using the left turn-signal is also appropriate when moving around an obstacle with the subsequent intention of turning right. Or suppose the road curves sharp right and immediately forks, so that (in terms of moving the steering wheel), we will either be turning right or hard right. Then the left turn-signal signals the former. Moreover, even when a car indicates and does move left, this will often be viewed as an incorrect use of the signal. Suppose that, coming up to a junction on a residential street, a car indicates left, but then turns abruptly with screeching tires and heads down a driveway. The left turn-signal seems to mean something like "I anticipate turning to the left, in the frame reference defined by my expected trajectory, at the nearest reasonable opportunity (i.e., no tire screeching)." But how then do we add in going around obstacles, overtaking vehicles or changing lanes? Or quickly 'pulling over' into a road shoulder for an emergency? Or the use of indicating to 'bag' or claim a parking space (i.e., that space on the left, that is about to be vacated, is 'mine')? Which uses should be included

[1] A sharp semantics-pragmatics distinction is built in much recent computational work in pragmatics, including the Rational Speech Act approach (Frank & Goodman, 2012, 2014) and the subsequent research tradition (e.g., Hu, Levy & Zaslavsky, 2021). This makes the approach difficult to apply directly to domains, as here, where no prior semantic conventions are agreed.

as putative of the conventionalized semantics, and which are matters of pragmatic inference is by no means clear; and whether a compelling semantics can be articulated is by no means clear.

The same issues arise equally, of course, in natural language, in which almost all words lack definitions (Fodor, Garrett, Walker & Parkes, 1980), seem best defined by a patchwork of interconnected uses (Wittgenstein, 1953), are modulated by capricious but metaphorical transformations (Lakoff, 1987), and the notoriously open-textured nature of language (Waismann, 1945), which legal scholars have argued makes contracts and laws inevitably incomplete and open-ended (Hart, 1961).

With these considerations in mind, we suggest that it is more natural to consider specific signals in concrete communicative situations as the primary vehicles of meaning; and to suppose that particular patterns of usage can become solidified into more stable conventions over time, as new communicative problems are solved by reference to similar prior problems.[2] But these conventions are never entirely stable, given the open-textured nature of language, and the continual arrival of unanticipated circumstances, for which some communicative solution must be improvised. From this perspective, there is no rigid distinction between fixed communicative conventions which are presumed to be part of the semantics and context-specific pragmatic inferences. Rather, individual communicative instances are primary; and broader semantic regularities emerge from a succession of such instances, each of which may shape those that come after (see Hawkins et al., 2021). This viewpoint is consistent with usage-based models of language structure across phonology, syntax and semantics (Bybee, 2006; Culicover, 1999; Goldberg, 2005; Langacker, 1987; Schmid, 2020), as well as language processing (Goldinger,

[2] Brennan and H. H. Clark's (1996) "historical" explanation of lexical "entrainment" exemplifies this perspective.

1988; van den Bosch & Daelemans, 2013), language acquisition (Ambridge, 2020; MacWhinney, 2014; Tomasello, 2003) and language change (Croft, 2000; Hopper & Traugott, 2003) and the interaction of all of these (Christiansen & Chater, 2016, 2022).

From this viewpoint then, we argue that conventions may, initially, be "instantaneous," only becoming increasingly entrenched after long periods of use; and, indeed, from this viewpoint, each new communicative use builds flexibly on previous uses, so that there will always be an element of in-the-fly interpretation, even of familiar linguistic signals (and, certainly, on any view, such flexible moment-by-moment reasoning will be required to resolve ambiguity, interpret referring expressions, including anaphora, and so on).

But does the very efficacy of such moment-by-moment reasoning in both the sender and receiver of a message, in alighting on the same mapping between signals and meaning, imply that the mapping is somehow "preferred" (whether because it is more efficient, easier to use, or links better with past experience)—and is not therefore wholly arbitrary. And is it not inherent in the very notion of a convention that it should be arbitrary—that there should be alternative, different conventions that would serve equally well?[3]

To clarify this point, it is useful to refer back to Lewis's (1969) classic study of conventions. For Lewis, what is crucial is that a convention is self-enforcing: each person should follow it, if everyone else does (in game-theory terms, this is a Nash equilibrium). But Lewis does not require that all conventions are equally "good." Lewis gives the example of a convention in his hometown of Oberlin, in the days when local phone calls were automatically cut off without warning after three minutes. When this happens, what are the speakers to do?

[3] We thank an anonymous reviewer for raising this point.

Assuming they would like to continue the conversation, one has to phone the other. But if they try to do so at the same time, the line will be blocked. The convention that arose was that the original caller would call back, and the other party would wait until they did. Apparently, newcomers to the town were informed about this convention on arrival. Lewis (1969) notes that "Other regularities might have done almost as well. It could have been the called party who always called back, or the alphabetically first, or even the older." (p.43) Note the crucial *almost*. The other conventions would not have been quite as good—the virtue of the prevalent convention is that it will always be the case that the caller has the other's number ready-to-hand, because they just called it. Any other convention will be slower and more awkward. Conventions based on the alphabet will also run into trouble with people who can be known by more than one name (e.g., *Bob* or *Robert*; or, if focusing on surnames, confusion over where *de Vries* sits in the alphabet). And age may not always be known, or even presumed, with potential social embarrassment. Nonetheless, whichever of these rules had been in force, any newcomer would benefit by following it—it would be self-reinforcing. So while there may be many conventions, some conventions are better than others.[4]

It is the very asymmetry between possible conventions that opens up the possibility that a convention—indeed, as we shall see, an entire system of conventions—can be created by a sender and receiver in a one-off communicative situation, without the need for prior communication. If both parties can realize that a particular convention is "best," then, potentially

[4] There are some conventions which seem genuinely arbitrary, such as driving on the left versus the right hand side of the road. These cases can arise due to symmetry considerations, but they seem to be the exception rather than the rule. For example, the mappings between phonology and orthography, between words and things, or merely the rules of grammar, seem paradigmatically conventional, though they have different properties regarding information-theoretic efficiency, perceptual, motor and cognitive processing requirements, ease of learnability, generalization and so on (e.g., Chater & Christiansen, 2010; Gibson et al., 2019; Morin, 2017; Sperber & Hirschfeld, 2004).

at least, they might simultaneously and independently be able to adopt that convention and apply it successfully. So instantaneous conventions are not a contradiction in terms. But for two parties successfully to alight on the same "best" system of conventions appears to require sophisticated reasoning. Below, we shall see that people do appear remarkably adept at agreeing instantaneous conventions, even when placed in abstract and unfamiliar communicative situations.

1.2. The challenge of multiple equilibria

Before a convention has been established through prior interaction, there are, as we have seen, many possible systems of communicative conventions that, if followed by both parties (we focus on dyadic interactions henceforth) would be self-reinforcing equilibria. Moreover, we have noted that not all such systems of conventions are equivalent—some will work better than others in the communicative context of the moment. If people are able to coordinate on the same convention 'in the moment,' then they must be able simultaneously to alight on the same convention.

In game-theoretic terms, the sender and receiver face a Schelling game, a special case of the wider class of coordination games (Clark, 1996; Schelling, 1960)[5]. In a Schelling game, people must simultaneously choose the same option as each other (e.g., given a collection of numbers, colors, dates, or locations, they are rewarded if they successfully choose the same

[5] In coordination games, more broadly, people typically need to choose compatible (rather than identical) moves to achieve mutual benefit. For example, one player might need to 'push' while the other 'pulls' in order to move a piece of furniture.

option). Here, the set of options on which coordination must be achieved is the set of possible mappings between signals and meanings.

In games of this kind, there are many equilibria—the games are defined so that, for any available option X, the case in which both players choose X is an equilibrium (that is, if one party selects X, then the other should select X too, and can only lose by choosing anything else); this is just another way of saying that the game requires that the players coordinate on the same choice. This structure clearly applies with communicative conventions, of course. The sender does not benefit from sending a message the receiver does not correctly interpret; and the receiver does not benefit by misinterpreting the sender's message. This applies even when the parties concerned have conflicting interests, or even outright hostility: one person may attempt to trick the other, for example, by sending a misleading message. But this hostile intent only works if the other successfully understands the message.[6]

The problem of simultaneously choosing a communicative convention has an interesting, and often underappreciated, consequence: that any account in which each individual is presumed to attempt to 'mind-read' the other (or otherwise predict the other's actions) cannot help with choosing between the multiple possible conventions. Thus, suppose A attempts to second-guess, and copy, whichever signal-meaning convention B has in mind; but A knows that B is similarly attempting to second-guess, and copy, whichever convention A has in mind, and so we are

[6] This raises the possibility of cases where the receiver of a message may actually benefit by failing to understand such a message (and hence fail to be tricked successfully). From the subjective point of view of the receiver, however, such misunderstanding will not seem beneficial, and the receiver will seek to understand the sender successfully. Roughly, it seems that when the receiver trusts the sender, understanding their message is plainly helpful; when they do not, understanding their message does no harm, because the receiver does not rely on its contents. A full discussion of these issues is beyond our scope here.

reasoning in a circle (for a broader discussion of circularity in social interaction, see Chater, Zeitoun & Melkonyan, in press).

In philosophy and linguistics, this problem of circularity is not faced directly. It is assumed that some literal meaning (semantics proper) can be taken for granted; and pragmatic elaboration of this literal meaning is assumed to be governed by fairly stable principles (whether conversational maxims, Grice, 1975; the principle of relevance, Sperber & Wilson, 1986; or generalized conversational implicatures, Levinson, 2000) which presumably are both known and applied by both sender and receiver. Hence, the question of how pragmatic departures from literal meaning are successfully coordinated by both the sender and receiver are not directly considered. Whether or not this approach is viable in the linguistic case (where the notion of literal meaning may be elusive, and the degree of pragmatic flexibility very great), this approach does not easily generalize to cases where signals have no conventional meaning (although see below).

Computational cognitive models of pragmatics can avoid circularity by assuming that linguistic conventions are already in place. They then typically assume that the behavior of one party (either the sender or receiver) can primarily be explained directly in terms of those conventions. The other party then has to adapt appropriately through some process of inference, perhaps including mind-reading. For example, Frank and Goodman's important model (2012) starts with literal meanings of words (defined by sets of objects to which they apply); and assumes that the sender chooses the word to describe an object in proportion to its specificity

(i.e., without reference to the mental state of the receiver).[7] The inferential, or mind-reading, part of the model is that the receiver uses Bayesian inference to infer the likely referent. It is not clear how this type of approach can apply here, as the signals are novel, and hence have no literal interpretation. It is possible to imagine an extension of this approach where some broader notion might take the place of literal meaning, but serving as the "default assumption" about the reference of a signal, even where no prior convention has been established. For example, in the context of our experiment described below, in which tokens are placed on "pads" which are spatially located near boxes which may contain positive or negative rewards, we could imagine a default assumption that placing a token near a box implies that this box should be selected. Then, pragmatic inferences could be built up, step by step. For example, we could imagine a set-up in which there are two boxes (call them Left and Right), and two available "pads" on which a token might be placed, which lie directly between these boxes. Suppose that both pads are closer to the Left box. Then the default meaning of placing a token on either pad would be that the Left box should be chosen. But then Frank and Goodman's analysis would be able to explain why using the rightmost of the two might actually come, pragmatically, to be interpreted as signaling the Right box (because if one wanted to signal the Left box, why not use the "clearest" left signal). It is an interesting question whether this style of approach can be developed for the types of experimental findings outlined below (e.g., where we will see the entire mapping between signals and meanings shifting, depending on the specifics of the communicative set-up). One challenge for the approach is whether a single notion of "default" meaning can be formulated for non-linguistic signals, in a way that applies across communicative situations. Thus, the same

[7] Goodman and Frank (2016) raise the concern of circularity, and how they side-step it: "…to avoid an infinite recursion and provide an entry point for conventional (semantic) meaning, the speaker is assumed to consider a simpler listener, the 'literal listener' P_{Lit}" (p. 820).

action of placing a specific token at a particular location might, in one context, seem naturally to signal that a nearby box should be chosen; in another, it might indicate the location of a hidden object; or signal an object that belongs to a particular individual. Moreover, in some contexts, the color, size or shape of the token, rather than its location, might carry the communicative load. It is not clear whether there is any equivalent default notion of meaning that is sufficiently stable to play the role of literal meaning in Frank and Goodman's approach. If we allow that a signal may have multiple "default" interpretations then we are, of course, faced with the challenge that both parties need to coordinate on the same default interpretation, which seems to lead us back into the very circularity with which we began.[8]

In the same way, conventional models of pragmatic inference, from Grice's (1975) conversational implicatures, to Sperber and Wilson's (1986) relevance theory, to Levinson's presumptive meanings (2000), take literal linguistic meaning as a starting point, and assume that either the sender or receiver, or perhaps both, elaborates upon this literal meaning in order to communicate. But these types of inference are not directly applicable here, at least without some extension as sketched above, because the novel signals used in our experiment have no literal meanings.[9]

How then can people alight on the same mapping between signals and meanings for novel conventions? It must be crucial, somehow, that all prospective conventions are not equal: some will be more useful, more efficient, less error-prone, more general, easier-to-use or perhaps

[8] Common ground is presupposed by each party, and used as the basis for simulating the likely agreed convention. But two parties may frequently differ in their beliefs about common ground—and indeed, having a common understanding of some aspect of the world is inherently a matter of degree. This raises the question of how the similarity of representations of the world can be understood and quantified. Interesting steps in this direction are taken by Feldman and Choi (2022).

[9] As noted above, there is also the possibility that literal meanings are actually a problematic notion, to be explained as emergent patterns from moment-by-moment communication.

easier to learn. One possible approach might be to pick one (or some amalgam) of these (or other) "virtues" of conventions, and to use this to rank conventions from worst to best; and to assume that people spontaneously choose the best available convention. If both sender and receiver have the same ranking of the goodness of conventions, then they will successfully alight on the same one, thus coordinating successfully without falling into circularity.[10] Moreover, they will alight on the best convention (according to whatever ranking is in play).

This approach seems a useful step forward in breaking the problem of circularity. But is there really a criterion of the goodness of a convention, that can be defined independently from the specific communicative challenge of the moment? The goodness of a convention, in a particular communicative context, will surely depend on the goals and constraints on the sender and receiver in that context—and these could be limitlessly varied. Not only that, the criteria for the goodness of a convention may differ for the sender and receiver. For example, compressed signals (perhaps with greater phonological erosion) might be more convenient for the sender, but more difficult to discriminate for the receiver. Or the sender and receiver may care about different aspects of the world, and hence prioritize the encoding of different meanings (e.g., perhaps one cares about shape and the other about subtle differences in color).

1.3. Solving the problem of multiple equilibria with virtual bargaining

We propose that a natural way to resolve both these problems—the vagueness of the goodness of a convention, and the potential different views between sender and receiver-—is by considering

[10] This approach could be formalized using a process of team reasoning (Bacharach, 2006; Colman & Gold, 2018; Sugden, 2003), which allows one of a set of multiple equilibria to be selected on the basis of some criterion of goodness.

that the preferred convention is the one that the sender and receiver would *agree*, were they able to discuss the best convention to use. Of course, the sender and receiver cannot actually discuss which convention to choose; but they can mentally simulate what agreement might be reached. These results of their mental simulation can only be based on their common ground. If either uses their private knowledge to inform the results of the simulation, then the other is unlikely to follow, thus compromising the central objective of coordination on the same communicative convention. The process of simulation can be viewed as a mechanism for establishing what Brennan and Clark (1996) call "conceptual pacts," concerning the communicative conventions of the moment—a mechanism which can operate even without interaction between individuals.[11] Thus, for example, at one moment *the cup* might pick out a mug among plates, but at the next, *the cup* might indicate a cup among mugs; and various phrases, such as *the little one*, *the cup*, *the blue cup*, *that blue thing*, etc. might most naturally refer to a particular, small blue teacup, depending on the other items available, the activity engaged in (e.g., drinking tea, matching crockery by their shade, etc.) and other contextual factors. Brennan and Clark note that such remapping is highly flexible and systematic, and we shall see below that the same is true for non-linguistic stimuli, where there is no prior literal meaning from which the process of forming a temporary agreement can begin.

Elsewhere, we have termed this process of mental simulation of a process of agreement "virtual bargaining," and the approach has been applied to experimental games (Misyak & Chater, 2014), driving (Chater, Misyak, Watson, Griffiths & Mouzakitis, 2018), models of social interaction (Chater, Misyak, Melkonyan & Zeitoun, 2016), moral psychology (Chater, Zeitoun &

[11] We would expect the same reasoning to operate when interaction is allowed, of course—as would be the case in explaining how small amounts of subtle communicative feedback allow people to create effective "pacts" to successfully coordinate (Brennan & Clark, 1996).

Melkonyan, 2019; Levine, Kleiman-Weiner, Chater, Cushman & Tenenbaum, 2018), and collusion in economic markets (Melkonyan, Zeitoun & Chater, 2018). More recently, virtual bargaining has been applied to problems of communication (Bundy, Philalithis & Li, 2021; Chater & Misyak, 2021), and we outline and extend this approach here.

Let us start by considering the experimental set-up in Misyak, Noguchi and Chater (2016), and on which the study below is a variation. As illustrated in Figure 2, suppose that there are three boxes, arranged in a triangle, and that it is common ground that each contains either a 'good' item (a *banana*) or a 'bad' item (a *scorpion*). One player knows the layout of scorpions and bananas, and can send signals to the other using a small number (in the experiment, one or two) of tokens, which can be placed on one or more of the boxes (depending on the number of tokens; i.e., one token per box). The second player observes the signals (i.e., where tokens are placed but not the associated movement trajectories of the sender) and can choose which box(es) to open. If a box with a banana is chosen, this is considered a positive outcome for both players; if a box with a scorpion is chosen, this is an (excessively large) negative outcome for both players. Hence, both players are motivated to coordinate on a convention for interpreting the placement of one or more tokens as indicating the layout of bananas and scorpions, so that the second player can obtain as many bananas as possible while avoiding the scorpions.

Suppose that it is common ground to the players that there is one banana and two scorpions, and that the first player (the sender) possesses one token; see Fig. 2. The "obvious" convention is to indicate the location of the box with the banana simply by placing a token on that box—and indeed, this is spontaneously and successfully used in our experiment. Yet, from a purely formal point of view, it would be just as effective to place the token on the box one step

clockwise (or one step counterclockwise) from the box with the banana. This strategy has a number of disadvantages, of course: it would be slower to implement (because each player needs

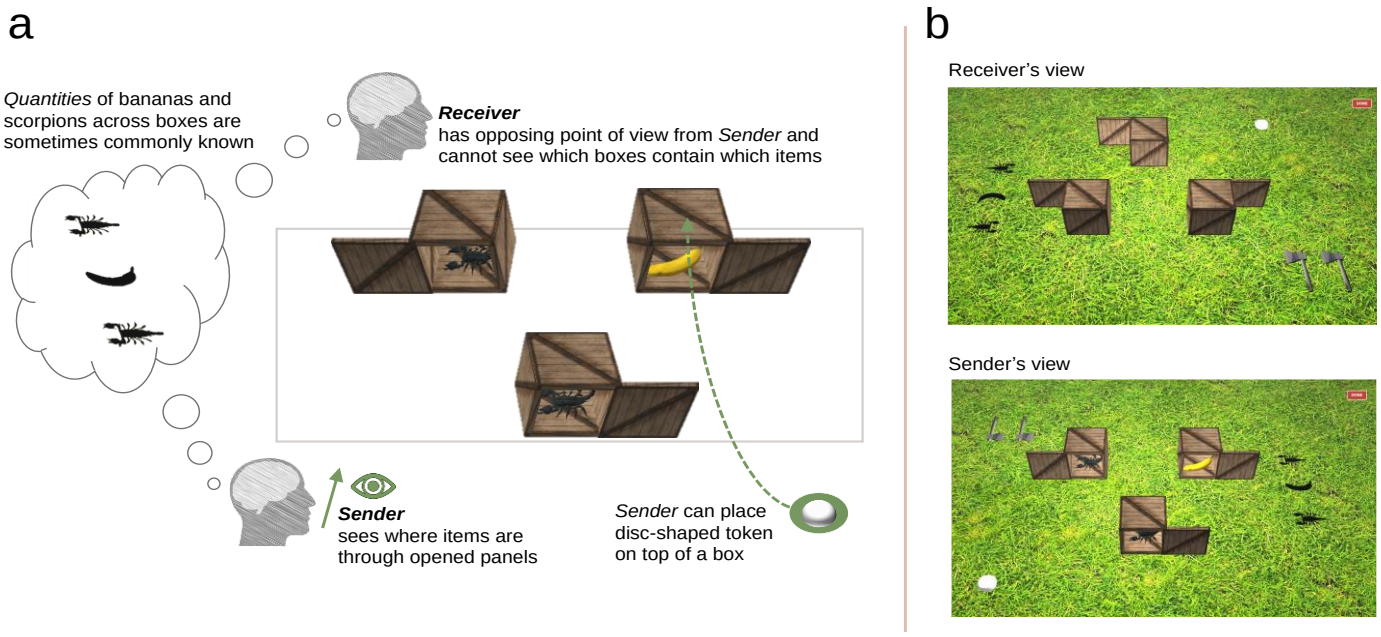

Fig. 2.

Simplified representation (a), focusing on key elements of one example scenario from the computer game in Misyak, Noguchi & Chater (2016). See the main text for discussion. Panel (b) shows the associated points of view of each player (sender and receiver), as would be observed by participants in the study; these images have been edited for small-print visibility in this illustration by enlarging objects in the scene and spacing them closer together. Crucially, quantities (but not locations) of bananas and scorpions are sometimes in players' common ground, as depicted by the shared thought bubble in panel (a); in the experiment, this is implemented by black 'imprints' of these items, corresponding in number, that can be mutually seen in the ground (see panel b). Other objects that can be perceived by both players as mutually viewable (here, for instance, one white token and two axes) are also part of common ground.

to make the relevant clockwise/counterclockwise adjustments), it is more likely to lead to mistakes (players may get mixed up between clockwise and counterclockwise), and it generalizes poorly (how does the strategy work if the boxes are no longer in a triangle, but are arranged in a line; or there are twenty boxes arranged randomly?). Yet another consideration is

that a ‘mark the box with the banana’ convention is unique, whereas the clockwise and counterclockwise alternatives are symmetrical, with no obvious virtue in choosing one over the other. The virtual bargaining approach does not aim to analyze or give weights to these factors in choosing a signal-meaning mapping in the specific communicative situation. It simply considers which convention would the sender and receiver choose, were they able to discuss and negotiate beforehand (i.e., where their challenge is: “if faced with a problem like this, how should we communicate?”).

To give an intuition for how the approach works, consider a variant on the situation above, where it is common ground that there are *two* bananas and one scorpion; and that the sender has one token. Here, people typically and spontaneously switch to a different convention: rather than ‘marking’ a banana with a token, they now mark the unique scorpion; and, realizing this, the receiver is able to choose the two unmarked boxes and thus obtain the reward of both bananas. Notice that the ‘meaning’ of the signal is reversed from trial-to-trial—from marking a box with a banana to marking a box with a scorpion. From a virtual bargaining viewpoint, this makes sense, because if faced with the hypothetical possibility of needing to communicate in this situation, both the sender and receiver can derive from their common knowledge that the ‘reverse’ convention will achieve a better outcome (yielding two bananas) than sticking with the original convention (yielding one banana). Thus, the sender and receiver are aiming to engage in what we might term “inference to the best convention,” to adapt a phrase from Harman (1965). But, crucially, such inference will only work if both come to the same conclusions. If one switches convention and the other does not, then they will fail to communicate successfully. According to the virtual bargaining account, the inference to the best convention proceeds by

inferring which convention would be chosen if the parties could negotiate—although any such inference can only be based on common ground, as no *actual* communication is possible.

Note that this approach is highly sensitive to details of the task that are in common ground. For example, suppose that in the two-bananas case above, it is also common ground that the receiver is only able to open one box. (In the Misyak, Noguchi & Chater, 2016, experiment, this constraint is implemented by reducing the number of axes from two to one, as each axe can only be used once to open a box.) Then, the 'reverse' signal, in which the box with the scorpion is marked is generally not used; instead the sender reverts to signaling one of the boxes with a banana, and the receiver chooses that box. Both the 'reverse' and 'standard' signaling conventions will, if adopted by both parties, yield a single banana. Yet it is possible for agreement to be reached by both parties reasoning that, could they communicate, they would agree something along the lines of "Why use this potentially confusing 'reverse' signaling when we don't have to? This could more easily lead to mistakes!" Because both parties can realize that the simpler convention would be agreed, then they can successfully coordinate on it. Thus, when it is commonly known that only one box can be opened in this situation, the simpler convention is typically coordinated upon by sender and receiver.

The notion of common ground is central to the virtual bargaining account (following Clark, 1996). As we have noted, using private knowledge to choose the "agreed" convention is likely to lead to miscoordination, because it will lead one party to draw inferences where the other cannot. Suppose, for example, that it is common ground that the sender has one token and the receiver can open up to two boxes. Suppose further, however, that whether there is one scorpion and two bananas, or the opposite, is not common ground. The sender observes the layout of the scorpions and bananas, and hence privately knows this information; but now this

information about quantities is not known to the receiver. (In Misyak, Noguchi & Chater, 2016, a set of 'imprints' on the ground matches the quantities of bananas and scorpions inside the boxes – see Fig. 2; but when the receiver's view of these imprints are blocked by a strategically positioned "wall," this information is no longer part of common ground). Now the sender should use the simple convention of marking a box with the banana, and the receiver should interpret this as such—the "reverse" convention would only be agreed if the parties' reasoning included common knowledge of there being a single scorpion and two bananas. (And accordingly, we found empirical support, as well, for participants' sensitivity to this change in common ground — albeit imperfectly [which may be due to the added difficulty of the sender's task on experimental "wall" trials, requiring differentiation between common ground and private knowledge; see, e.g., Hanna, Tanenhaus, & Trueswell, 2003; Keysar, Bar, Balin, & Brauner, 2000].)

Note, finally, that if bargaining is really a crucial factor to the convention chosen, then this yields the distinctive prediction that the bargaining power of the sender and receiver should have the potential to change the signal-mapping convention used. Consider the following thought experiment.[12] The following set-up, depicted in Figure 3a, is common ground. There are two sets of three boxes. In one set, the boxes have rewards ($8, $2), ($0, $0), ($0, $0) (where the first number in each pair is the reward for the sender, and the second for the receiver). The other set of boxes have rewards ($1, $4), ($1, $4), ($0, $0). The sender can place a single token on one of the six boxes. The receiver can open no more than 2 boxes. The sender, furthermore, has private

[12] The discussion below goes beyond prior discussion of the virtual bargaining approach to communication (e.g., Chater & Misyak, 2021).

knowledge of which set of prizes have been allocated to which set of boxes, as well as the allocation of specific prizes within a box set (as, e.g., in Figure 3b).

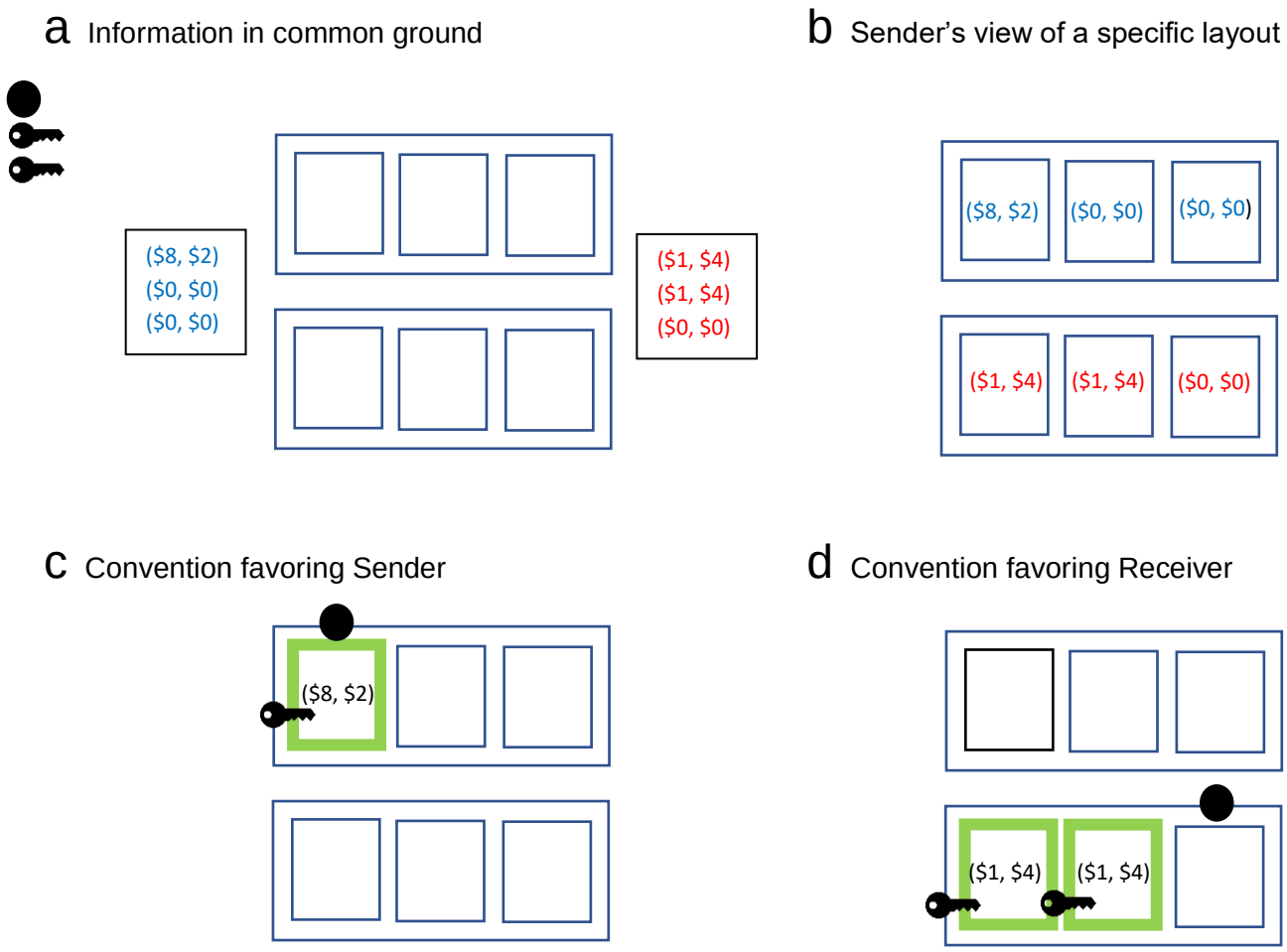

Fig. 3.

Thought experiment illustrating the impact of bargaining power on the chosen convention. Panel (a) illustrates the common ground for the sender and receiver. There are two sets of three boxes, and two sets of three prizes (blue, red). Whether the blue or red prizes are allocated to the upper or lower set of boxes is not commonly known; neither is the location of the prizes within each set of three boxes. The sender has one token (represented by the filled black circle); the receiver can open up to two boxes (with the two black keys). Panel (b) shows a specific allocation of prizes to boxes, which is seen only by the sender. The convention favoring the sender is as follows (c): the sender places the token on the square which is most favorable to her, and the receiver opens it with payoffs ($8, $2). The convention most favoring the receiver is as follows (d): the sender indicates the unique ($0, $0) prize among the group of three and the receiver opens the other two boxes, with two payoffs of ($1, $4), and a total payoff of ($2, $8).

Now imagine that the sender and receiver could actually bargain, to decide how they would communicate in a case like this. As shown in Figure 3c, one possible convention, C_S, would give the sender \$8 and the receiver \$2: the agreement would be that the sender would mark the (\$8, \$2) box, and the receiver would open it. As shown in Figure 3d, another possible convention, C_R, (using a 'reverse' mapping) would give the sender \$2 and the receiver \$8. This convention is that the sender marks the (\$0, \$0) box which is unique in its set of three boxes; and the receiver opens the other two boxes, each with payoff (\$1, \$4).

Whether the players coordinate on conventions C_R or C_S (favoring receiver or sender) will depend on any number of factors. For instance, the chosen convention may depend on the relationship between the sender (S) and receiver (R). Suppose that it is common ground that S is rich and benevolent, and R is poor and desperate for money. Then it is common ground that the "best" convention is C_R. Thus, the marking of the box indicates the (\$0, \$0) box that is unique among the three prizes. By contrast, suppose S is poor and desperate for money, and R is rich and benevolent. Then, the marking of the box will indicate the (\$8, \$2) option and R will choose this, giving S a more positive outcome.[13] Or suppose, for example, that one party is a mafia boss and the other is a side-kick; then the players will likely coordinate on whichever convention benefits the boss (indeed, for the side-kick to make the opposite choice would likely be viewed as presumptuous). Or suppose the players have just answered a quiz as a team, to win the opportunity to play this game. If one player answered all the quiz questions and the other knew

[13] We will ignore the space of equivalent, but more awkward, conventions, such as indicating a particular box by marking the box immediately to the left, with wrap-around. For the reasons stated above, these would be rejected by bargainers and needlessly complex and likely to produce mistakes. We also ignore the many conventions with strictly inferior outcomes (e.g., S indicates a box, and R randomly chooses two of the boxes from the other set).

none of the answers, then it seems likely that both would coordinate on whichever convention favors the more knowledgeable player, perhaps as a matter of equity (see, e.g., Berkowitz & Walster, 1976; Hoffman & Spitzer, 1985).

In each of these scenarios, a natural way to understand which communicative convention people coordinate on (and hence, the meaning associated with the placing of the token), is that it depends on what they would agree, were they able to negotiate; and based on common ground between the players, the results of this discussion can be imagined, and hence acted upon, even without communication actually being possible. So, for example, the rich, benevolent person may imagine saying that the other person has greater need for the money, and the other gratefully accepting the relevant convention. Similarly, the side-kick can imagine all too easily the barked instructions from the mafia boss, which brook no dissent. And in the quiz example, both players might imagine the less knowledgeable one conceding that the other deserves most of the reward, having got them successfully through the quiz. In short, coordination works when both parties can imagine what agreement would be reached concerning the 'appropriate' convention, given their common ground. In many circumstances, of course, the imagined negotiation may not have a mutually "obvious" outcome (e.g., the rich player may not also be benevolent), and in such cases, coordination on the same convention may fail, with a consequent breakdown of communication.

According to the virtual bargaining account, then, there is no fixed criterion of which counts as the "best" communicative convention (in terms of efficiency, generalizability and so on); rather what counts as the best convention will depend on whatever factors can impact the imagined process of negotiation between sender and receiver. And, as in the Misyak, Noguchi and Chater (2016) studies, the convention used can flexibly shift from one trial to the next. We

shall see shortly that signal-meaning mappings can also "instantaneously" remap an entire system of conventions.

2. Forming instantaneous systems of conventions

Coordinating on the right signal-meaning mapping seems to require overcoming formidable cognitive challenges. As we have indicated, to establish the meaning of any particular signal, people need to establish the wider signal-meaning mapping of which that signal is a part. Moreover, as we have noted, to communicative successfully, they need to choose the *same* systematic convention, in the expectation that that convention is the one they are most likely to have both alighted on. One plausible suggestion, which we will explore below, is that the chosen system of conventions should be, by some criterion, optimized to the demands of the situation.

This appears necessary to explain the flexible understanding of signals in driving in Figure 1. We note that the specific details of the road layout shift the interpretation not merely of the individual signals of indicating left, right, or not indicating at all, but also shift these systematically, so that each of the three signals has a distinct meaning. But our interpretation of driving signals does not necessarily arise 'in the moment,' but might be heavily influenced by individual learning and cultural evolution over large numbers of previous driving interactions. Indeed, the on-line reasoning required for the sender and receiver to flexibly coordinate on a system of communicative conventions appropriate for the specific situation in hand is likely to be highly sophisticated. With these considerations in mind, it is natural to be skeptical that entire systems of conventions can flexibly be generated on-the-fly, in novel situations. The results

reported below show, though, that people are indeed able to create such systems of conventions in the moment.

2.1. Relationship to prior experimental work

The approach outlined below has commonalities with a number of research traditions, but is distinct from each. First, there is, as we have noted, an important tradition of work exploring how people flexibly adapt existing communicative conventions, and particularly natural language, when dealing with specific, and often novel, communicative challenges (e.g., using natural language to describe unfamiliar tangram figures, Clark & Wilkes-Gibbs, 1986). Experimental pragmatics uses laboratory studies to investigate pragmatic inference and constraints in our everyday use of established language, examining phenomena such as reference resolution, implicature, common ground and spoken dialogue (e.g., Brennan & H. H. Clark, 1996; Degen & Tanenhaus, 2015; Garrod & Anderson, 1987; Hanna et al., 2003; Hawkins, Frank & Goodman, 2020; Horton & Keysar, 1996; for an overview, see Noveck & Reboul, 2008; see also Noveck & Sperber, 2004). Here, by contrast, we study core pragmatic principles of human language and communication, but during the *de novo* emergence of minimal communicative systems under controlled settings.

A more recent research tradition in experimental semiotics (e.g., De Ruiter et al., 2010; Galantucci, 2005; Healey et al., 2002, 2007; Roberts & Galantucci, 2012; for a primer, see Galantucci, Garrod & Roberts, 2012), by contrast, usually does focus on the emergence of de novo communicative conventions in the laboratory. Typically, these conventions emerge, often relatively rapidly, through successive interactions between experimental participants, such as in

conveying a location or layout from one participant to another through the movements of 'agents' in a computer game (e.g., Scott-Phillips, Kirby & Ritchie, 2009), or in coordinating actions with communicative signals (e.g., each person controls an 'agent' and the goal of their interaction, through the use of a novel graphical medium, is for the agents to meet at the same location on a grid, Galantucci, 2005). Sometimes, in Pictionary-style communication tasks, iconic representations are used as the starting point for iteratively deriving new graphical signals through participants' repeated interactions (e.g., Garrod et al., 2007). Partner feedback and interactivity can be crucial for success in these tasks, and allows for the cumulative shaping of conventions; in our work reported below, by contrast, participants must communicate in 'one-shot' encounters without such rich feedback dynamics.

Finally, note that there is a related, and overlapping, line of research which focuses on the way in which communicative conventions across a community of agents is shaped through multiple interactions, whether sequentially or involving a more complex network of interactions (e.g., Hawkins, Franke, Frank, Smith, Griffiths & Goodman, 2021; Kirby, Cornish & Smith, 2008; Kirby, Tamariz, Cornish & Smith, 2015).[14] Frequently, in many behavioral studies, participants begin these experimental tasks 'initialized' with pre-specified (and random) mappings, which are then subsequently shaped or structured into new mapping conventions by participants. For example, Winters, Kirby and Smith (2018) find that, when developing conventions by which novel signals (artificial "words") map to cartoon "aliens," the degree to which the alien referents are predictable affects the brevity and contextual flexibility of the conventions established. Relatedly, Kanwal, Smith, Culbertson and Kirby (2017) show that the

[14] Not all studies involve, methodologically, *communicative interactions* as such between participants; but may instead use, exclusively, diffusion chains of individual participants to investigate effects of cultural transmission (e.g., Kirby et al., 2008).

mappings between artificial "words" and objects are shaped by the specific communicative demands of the experimental set-up (and, under plausible conditions, gives the tendency to associate brevity and frequency of use exemplified in Zipf's Law in natural language [Zipf, 1935]). The cultural evolution of communicative conventions has also been extensively explored using interactions between artificial agents (e.g., Brochhagen, 2020; Hu, Levy & Zaslavsky, 2021; Peloquin, Goodman & Frank, 2019; Spike, Stadler, Kirby & Smith, 2017; Steels & Belpaeme, 2005; Steels & Loetzsch, 2012).

The focus here is more elementary than the focus of these research themes, and might provide useful foundational assumptions for these more complex cases. In particular, we minimize experience of communicative conventions, by using an abstract communicative task, rather than exploring the flexible use of existing conventions (in distinction from experimental pragmatics). We focus on the flexible creation of conventions in one-shot interactions with (ostensibly new[15]) communicative partners (rather than allowing the rich interactional dynamics in experimental semiotics). Finally, our primary interest is how conventions may arise instantaneously, and only secondarily the implications for the conventions eventually adopted by the community of people who can communicate (in distinction from work on the cultural evolution of language). An important objective for future work is to consider how the results outlined here may feed into work in these three traditions.

Our use of an abstract task allows us to experimentally vary available meanings (the 'turnings' in our motoring example become possible locations of 'target' object(s)), in order to test whether people can create and systematically reorganize their communicative conventions

[15] See Section 2.2.4 for further detail.

on the fly. We also ask whether modifying the available set of signals (equivalent to adding honking or high beam flashing, or removing one turn-signal from the vehicle, in our motoring example) can cause people to modify the entire signal-meaning mapping. In both cases, we find that people are adept at such remapping. We find preliminary evidence, moreover, that people reorganize mappings only when confronted by a new communicative context for which the current mapping is inadequate. This latter point has interesting implications for the cultural evolution of communicative conventions, suggesting that people will tend to adhere with an established system of conventions in a new context where it performs adequately (though it may not otherwise be the most natural convention for that new situation). In such cases, conventions will become entrenched. But entrenchment appears to be provisional: where demanded by the new communicative context, people can spontaneously abandon an established signal-meaning mapping and adopt another mapping, better tailored to the communicative demands of the moment.

2.2. Experimental Study: Method

2.2.1. Participants.

We recruited participants for two signaling conditions ("3:4" and "4:3;" described below), with each condition comprising four sessions of different participant groups. Recruits were University of Warwick undergraduates, with self-reported good (or excellent) fluency in English and unimpaired color vision.

Twenty-six spaces were advertised per session, the maximum number of networked computers that could be reliably accommodated for our task's design and software. Anticipating

potential no-shows, 8 sessions were conducted in order to reach a total of approximately 160 participants, contributing 40 dyads (sender-receiver pairs) per condition. We projected 40 dyads to be sufficient based on related pilot data from in-group interactions (with different partners), which was underpowered (due to resource limitations on maximum group size). The possibility of network-level effects also led us to plan to simulate the outcomes of multiple communicative "communities" through many fixed-partner dyadic interactions.

Recruits were informed, prior to signing up, that sessions required an even-number of participants and, as such, that one person might be dismissed at a session's start with £3 payment. Of the 203 sign-ups, 42 recruits did not attend across the eight sessions. Three sessions had an odd-number of attendees, so three people were randomly selected for dismissal.

Each of the remaining 158 students gave informed consent and participated in one of the two conditions, forming 40 dyads in total for the 3:4-condition sessions (12, 10, 9, and 9 dyads) and 39 dyads in total for the 4:3-condition sessions (9, 10, 10, and 10 dyads). Ten pounds was paid to all participants (52 male, 104 female, 1 trans female, 1 undisclosed; mean age = 19.9 years, *SD* = 1.4, with 2 participants of unknown age).

2.2.2. Communicative Task.

We developed a new variant of a computer-based communication game used in Misyak, Noguchi & Chater (2016). A sender and receiver must coordinate, by sending and interpreting minimal signals, in order to collect target objects 'hidden' in a scene. Partners view the scene from opposite visual perspectives, affording them with asymmetric knowledge of target locations and role-specific access to tools (a signaling-indicator for the sender, an object-selector for the receiver).

Each scene included two boxes, which were horizontally flanked by a set of four equally spaced pads (circular, flat spots; see Fig. 4). The positioning of the pads and boxes was held constant, such that: the left-inner pad (second pad, from left to right) was directly adjacent to the left-side box, with the leftmost pad following next in terms of proximity to the left box; and the right-inner and rightmost pads were aligned analogously with respect to the right-side box. The content of each box—either a banana (target) or a scorpion (foil)—varied on each trial and was visible only to the sender (by means of box panels that automatically slid open on the sender's side). However, only the receiver could choose boxes on behalf of the pair. Across trials, the joint goal was for partners to coordinate on choosing all, and only, the boxes containing bananas.

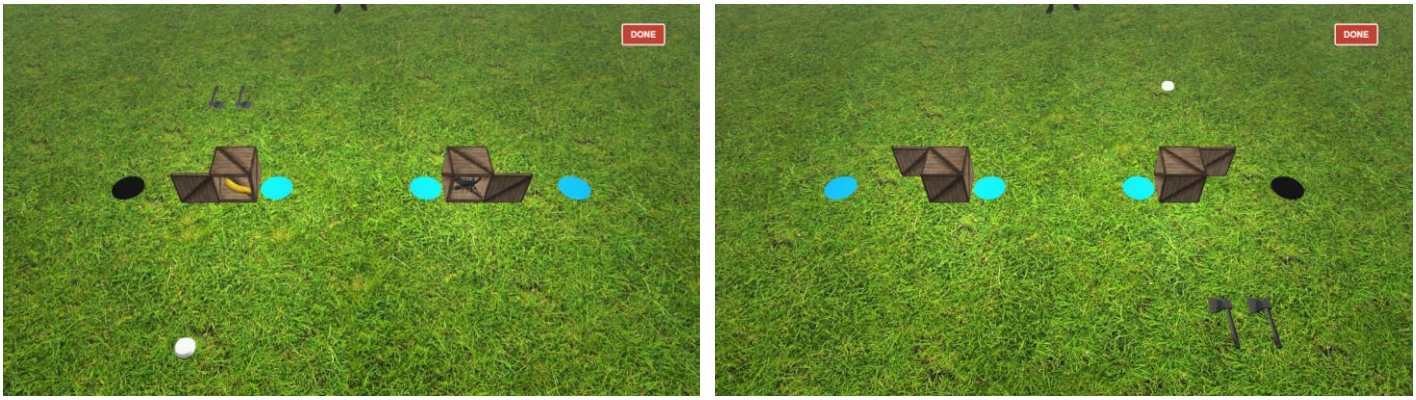

Fig. 4.

A scene example from the communication task, showing the different visual perspectives of sender and receiver. Each scene contained two boxes—with a banana or scorpion inside, and visible only to the sender. The color (blue, black) of signaling pads varied on each trial. The sender always had only one token available (bottom-left of Sender's view) to potentially place on any one of the blue pads, in order to convey the location of any banana(s) to the receiver. The receiver then selected up to one box per provided axe (shown bottom-right, Receiver's view). A red "DONE" button appeared at the scene's upper-right. See the Method for more information about these elements, and how box contents and pad colors varied across trials.

To do this, the sender could first transmit a signal by moving the solitary ‘token’ (white disc-shaped object) onto one of the ‘available’ signaling-pads. Whereas the four pads’ locations in the scene were always the same, their colors varied from trial to trial: a token could be placed on top of a blue pad; black pads were ‘unavailable,’ meaning that tokens could not be placed on them. Using the computer-mouse to move the token onto one of the blue pads, the sender clicked the “Done” button, which updated the receiver’s view of the scene with the signal of the token atop the pad.

The receiver then used the computer-mouse to move the axe-object on top of the box to be selected. Depending on task phase (described below), the receiver sometimes had a second axe available, so that a second box could also be selected. When finished placing axe(s), the receiver clicked the “Done” button. Any selected boxes were then “axed” open, revealing their contents to both partners. Additionally, the side panels of any unchosen boxes were slid open on the receiver’s side to convey full feedback to both partners on the successfulness of their interaction (i.e., whether they collected all and only the bananas).

Senders may find it useful to transmit a “null” signal under certain circumstances (e.g., when none of the boxes contained targets); and receivers need not select any box (by simply not performing any actions before clicking “Done”). When this happened by the sender and/or receiver, the text-message “Your partner did not place anything” displayed on their partner’s screen, and the trial-sequence continued per the above steps.

2.2.3. Design.

To collect bananas successfully, both sender and receiver need a common strategy for accurately

linking the ‘indicated’ pads (signals) with the specific locations of targets (meanings). Of key interest is whether these nascent mappings of meanings to signals are created independently, or interdependently as part of a system. If participants create communicative conventions in isolation, then a signal’s meaning should be interpretable on its own. However, if participants construct conventions as part of a system, then—by changing the set of signals or meanings—the interpretation of the elements of the whole-system can be shifted. We therefore varied both the sets of available signals and meanings to be expressed.

Signaling options were manipulated by restricting the availability of individual pads across trials, resulting in four configurations of accessible and inaccessible pads (see Fig. 5a). In one arrangement, all the pads were blue, producing a set of five signaling options (each of the four potential pad placements for the token, plus the always-available 'null' signal of not moving the token to 'indicate'). Two other arrangements resulted from taking each of the outermost pads, in turn, out of potential use (as represented to partners by switching the singly affected pad's color to black). In the fourth arrangement, both outer pads were unavailable, leaving only the two inner pads available.

Meanings were manipulated by changing the number of unique scene “events” (i.e., the possibilities for target locations) for the sender to communicate. One set of events consisted of the full complement of permutations for locations corresponding to two, one or no bananas (as depicted in Fig. 5a, right side). The other set was identical except for the removal of two-banana occurrences, thereby containing exactly three distinct events (as depicted in Fig. 5a, left side).

Senders had ultimately to convey the full range of meanings corresponding to a given event set, over each signaling-pad set, across many trials. For any particular trial, one of the four

signaling-pad configurations would be instantiated and one of the scene events (drawn from an event set) would need to be communicated, by choosing whether and where to place the token among the available pads. In the first half of the experiment ("early period"), participant-dyads were exposed only to events drawn from one of the event-sets, based on experimental condition. Subsequently, dyads were switched to the other event-set in the "late period," or second half of the experiment. Participants were made aware, only just prior to each period, of the number (either 3 or 4) and nature of events they might encounter. This resulted in two separate conditions—3:4 and 4:3, referring to the number of events defining the early:late periods. (E.g., in the 3:4 condition, participants completed the 3-event phase [Fig. 5a, left], followed by a 4-event phase [Fig. 5a, right].)

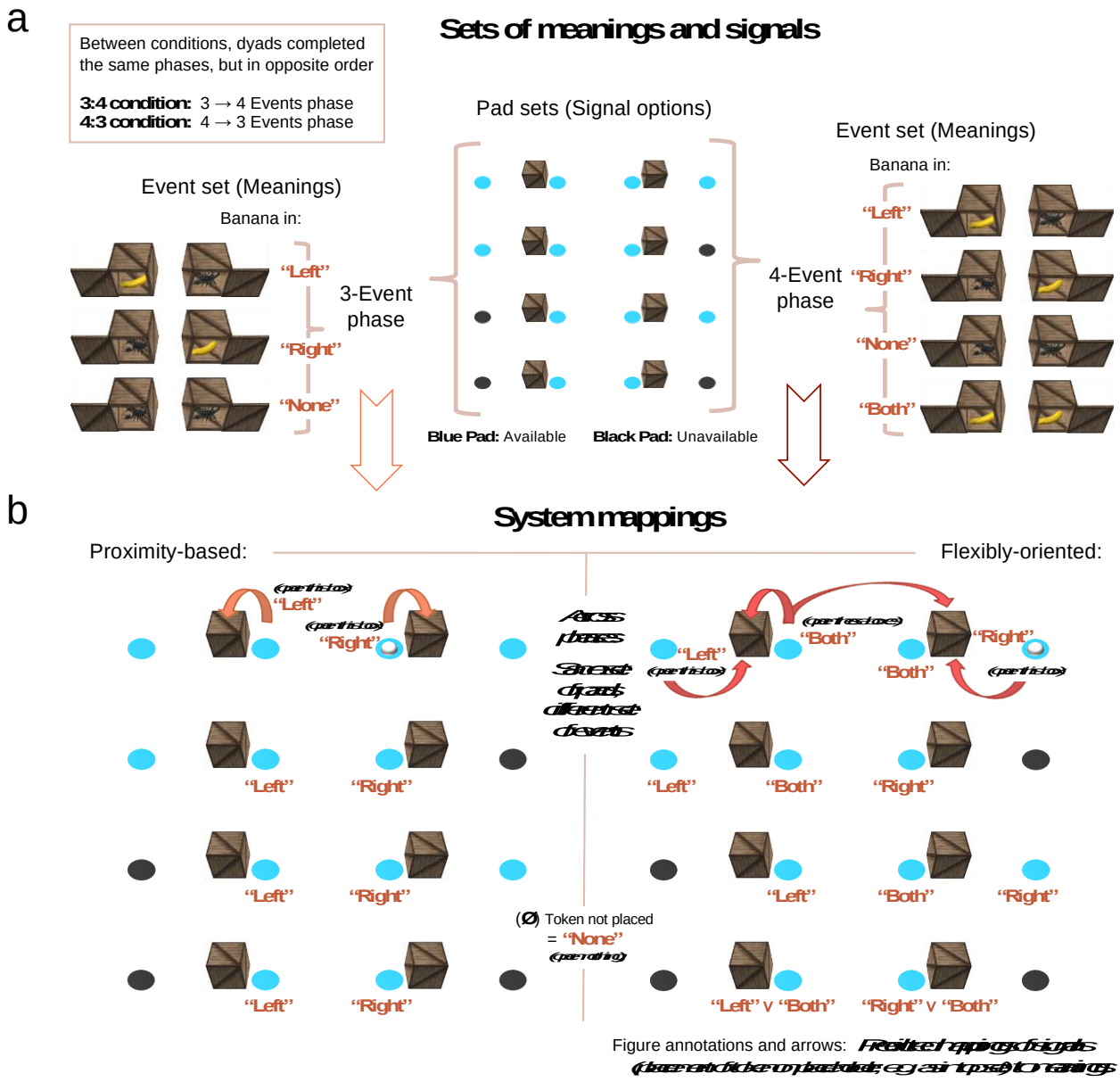

Fig. 5.

Hypothesized system-level mappings, as a function of the sets of signals and meanings available to dyads in each event phase. There were three possible meanings (target locations) to convey (a, left side) in the 3-event phase, and four (a, right side) in the 4-event phase. In both event-phases, there were (a, center) four distinct pad configurations, affording different sets of signal options. Blue pads were available for signaling; black pads were not. In the early period of dyads' communication, (b) proximity-based and flexibly-oriented mappings of signals-to-meanings are expected in the 3-event and 4-event phase, respectively. Under these system mappings, the interpretation of a potential signal (the placement of the token on a given pad; e.g., as illustrated in the top pad set) is indicated by the shorthand annotation next to the pad (e.g., "Left" = Banana in left box). (Box panels in the experimental task would have appeared open to the sender; they

are depicted as closed for expository purposes.)

Crossing each scene event with each of the four signaling-pad configurations yielded 12 unique trial-types in the 3-event phase; these same trial-types, plus an additional four trial-types (corresponding to the double-bananas case), made up 16 trial-types in the 4-event phase. For comparison between conditions, we used the same presentation-order of trial-types (within an event-phase) for all experimental sessions. Specifically, for the 4-event phase, one random trial-sequence was generated, which presented each of the 16 trial-types once per block, for two blocks. A similar procedure was followed for the two blocks of the 3-event phase, with the exception that would-be instances of the double-banana event were replaced with the double-scorpion event; this replacement provided exposure to an equal amount of scenes as in the 4-event phase. When these two sequences were ordered together for each condition, they also obeyed the constraint of avoiding consecutive repeats of trial-types at the transition between periods.

Signaling options therefore varied trial-to-trial across the full experiment, while the set of meanings varied both between-subjects (as in the early periods of the two conditions) and within-subjects (from the early to late period).

We reasoned that, when considered in isolation, the most reasonable signal to indicate the location of a sole banana would be based on proximity: a token would be placed on either the left-inner pad if the banana is in the left box, or the right-inner pad if the banana is in the right box. Also considered independently, we reasoned that the most reasonable signal for no bananas would be the "null" signal, not placing the token anywhere. When these possibilities (banana in the box at "left," "right," or "none") compose the exact range of meanings to be conveyed and the inner pads are available (as both satisfied in the 3-events phase), these 3 signal-meaning

mappings also function well congruently and contrastively as a system of mappings. Therefore, on either basis of convention formation, we would expect participants in the 3-event early period to simply implement these proximity-based mappings (as illustrated in Fig. 5b, left).

However, when participants need a way to express "both" (as in the 4-event phase), the most reasonable signal for "both" (again, derived in isolation) would likely be either of the inner pads (as they are intermediate between both boxes, and thus proximal to each). Such a mapping, though, conflicts with the independently-best mappings for "left" or "right" (assuming the rationale from above); so if participants adhere to constructing conventions independently, this should lead to problems for successful interpretation.

Conversely, if participants construct mappings as part of a system, then they should link meanings and signals with contrast to other possible meanings and signals, such that they interdependently differentiate (to the extent possible) each of the events to be communicated. One apt way to do so for the 4-event phase would be to use the location of available pads within the pad-set (or equivalently, the relative positioning of blue pads to the two boxes) – leftmost, medial, or rightmost – to respectively signal that the target(s) are located in the left-only box, both boxes, or right-only box. If this is so, then the interpretation of any one signal should be capable of shifting, in relation to other possible signals, as the pad-set changes.

This "flexibly-oriented" strategy for forming mappings is illustrated in Fig. 5b, right, for the various signaling-pad configurations. When the full set of signals are available (as in the top pad-set), either inner pad can be used to signal the double-bananas event, whereas the outer-left and outer-right pads can be used to signal the location of a sole banana in the left or right box, respectively. However, when the outer-right pad is unavailable (second pad-set from the top), the

signal for the banana in the right box shifts to the inner-right pad, which is now no longer the signal for "both." Similarly, when the outer-left pad is unavailable (third pad-set from the top), the signal for the banana in the left-box becomes the inner-left pad, whereas the inner-right pad – being in the middle of the available pads – now becomes exclusively mapped onto indicating both boxes. Finally, for the bottom pad-set with no outer pads available, each inner-pad indicates the presence of a banana in at least the adjacently-situated box, if not both boxes.

Notice, then, that under this flexibly-oriented strategy, the interpretation of any one pad's meaning as a signal depends not solely on the signal itself, but on the set of other signals that could potentially be used; and thus interpretations shift accordingly as the set of signals changes. Additionally, if we see participants adopt this strategy in the early 4-event phase and the proximity-based strategy in the early 3-event phase, this would indicate that the interpretations of signals also depend on other meanings that could be expressed, and is thus systematically reordered when the set of meanings changes. Both kinds of entire re-mappings of meanings-to-signals would be expected if people construct systems of communicative conventions (and do so flexibly and instantaneously from the onset), but would not be predicted from an approach to interpretation which attempts to assign a meaning to a signal independently of the other signals and meanings available.

But what happens when the set of meanings changes across time (i.e., between the early and late periods of our experiment)? One possibility is that early communicative conventions become entrenched and transfer to the later period to which they were not adapted. Another possibility is that dyads create communicative conventions anew, as if optimized to match the constraints of the new period. A third possibility is that conventions become transferred unless they are inadequate for the new context, in which case they flexibly reorganize. If so, then the

entrenchment of a convention might be provisional, so that it can spontaneously be overturned when it is inadequate to new communicative demands, and replaced by an alternative, more appropriate, system of conventions. We have no prior predictions among these possibilities, but explore this additional issue through our experiment.

2.2.4. Procedure.

Participants completed the communicative task over networked computers during their group session. Within each session, we anonymously and randomly sorted participants into fixed-partner pairings (sender-receiver dyads) using mild deception to nonetheless simulate different-partner interactions. Specifically, participants were instructed that they would be matched with different people on different trials throughout the task (and only debriefed afterwards that they had been matched with the same partner). Consistent with this impression, advancement of partners to the next trial was group-paced (to ostensibly enable new partner-matchings), rather than synced to the timing of individual pairs.

Participants were also instructed on other aspects of the game, including the sender-receiver turn-taking structure, procedural details (e.g., the manipulation of tokens and axes), scene details (e.g., that blue pads are accessible for token placement), and the joint goal of collecting bananas (clarifying that success in the task means collecting the *exact* number of bananas that are present and no scorpions). Initial instructions also outlined the possible events (3 or 4); this was specific to the early event-phase of their condition, without referring to any later period or subsequent change in the range of events.

Prior to the task, participants performed 4 practice trials, corresponding to two scene events: one consisting of a banana in the left box only and the other with a banana in the right

box only. The number of axes in each scene was congruent with the early period of experimental trials (2 axes in the 4:3 condition, and 1 axe in the 3:4 condition). However, the procedure differed from experimental trials in two respects: 1) each scene event was presented twice successively (as two trials), with participant-roles reversed on second presentation to familiarize participants with the differing visual perspectives; and 2) there were two blue signaling-pads situated on top of the boxes (rather than four pads on the ground) to avoid biasing convention-formation before the experimental trials.

Participant-pairs completed the experimental task, consisting of 64 experimental trials across 4 blocks. A person's role as sender or receiver remained constant within a block, switching after each block. At the transition between the second and third block (marking the division between event-phases), participants received further instructions that outlined the imminent change in the number of possible events (and axes available) and were reminded to be as accurate as possible.

After the task (but before debriefing), we administered a questionnaire, which probed the effectiveness of the partner-matching deception. Participants were told, “For some groups when we run these sessions, we match a participant with the **same partner** on every trial (regardless of the task instructions). And for some groups, we match a participant with **different partners** on different trials.” Participants were asked if they thought it was possible to tell which of these groups they were in [Yes / No], and then prompted to make a guess [Same partner / Different partners].

2.3. Results

Study data can be found at the OSF repository: osf.io/t58ys. Questionnaire analyses confirmed the effectiveness of the partner-matching deception to simulate different-partner interactions. Roughly half of participants (52.5%) responded that they could not tell whether they had been matched with same or different partners. Among the other half, nominally more participants in the 4:3 condition (55.1%) than in the 3:4 condition (40.0%) answered affirmatively that they could distinguish the matching; however, this difference was not statistically significant: $\chi^2(1) = 3.63$, $p = .057$. More pertinently, there was no association between self-reported ability to tell whether partner matching was fixed and the accuracy of participants' guesses in this regard: $\chi^2(1) = 0.66$, $p = .416$. Indeed, participants' guesses were no better than chance-level—both overall (with 51.3% of all participants guessing different partners, 48.7% guessing same partners: $\chi^2(1) = 0.10$, $p = .750$) and specifically among the subset who thought that they could tell (with 54.7% guessing different partners, 45.3% guessing same: $\chi^2(1) = 0.65$, $p = .419$).

For our main analyses, we focused on the signals generated by dyads, treating each dyad as a unit. As there were no group dependencies (per design) and sessions were identical (same trial-presentation order within condition), data was collapsed across sessions for each of the two conditions. No data observations were excluded.

2.3.1. Changing the set of meanings between conditions and periods.

To examine the effects of changing the meaning-set, we identified signals for key trial-types common to both event-phases, in which dyads' use of flexibly-oriented or proximity-based mappings could be distinguished from each other and directly compared. These were cases in which both an "inner" and "outer" pad were available as choices for signaling a single target in the box between them, as outlined in Figure 6a. There were two such cases for when the target was in the left box and two cases for when the target was in the right box.

For these key communicative cases, we accordingly computed the proportions of sender signals in which a token was placed on the outer and inner pads nearest to the target box. Per Figure 5b, token placement on the outer pad in these cases fits the flexibly-oriented mapping, whereas token placement on the inner pad in these cases fits the proximity-based mapping. Although other signals were permitted (i.e., no token placement, or placement on a more remote pad), only one such instance of an alternate signal (< 0.5% of key trials) occurred across conditions.

Figure 6b shows the group-averaged proportions of inner and outer pad signals that dyads transmitted per left-box and right-box target, and separately for each event-phase. A clear majority of inner pad signals, at the expense of outer pad signals, characterized the 3-event phase (early period) of the 3:4 condition; this pattern appeared almost precisely inverted in the 4-event phase (early period) of the 4:3 condition, in which outer pad signals predominated over inner pad signals. In the later periods, there was a strong preponderance of outer pad signals in either condition.

Signal proportions were largely symmetrical for left-box and right-box targets, and therefore collapsed across for statistical tests. We first compared initial differences in signaling when the meaning-set had 3 versus 4 events, as occurred between conditions in the early period. As expected, the proportion of outer pad signals (which reflects use of the flexibly-oriented mapping) was significantly greater for dyads experiencing the 4-event set in the 4:3 condition (*Mdn* = 1.00, 96% confidence interval[16], or CI = [.75, 1.00]) than for dyads experiencing the 3-

[16] 95% confidence intervals for medians would be approximations, due to the discreteness of the binomial distribution. CIs reported for medians throughout provide an actual coverage of 96.15%. Significance values for nonparametric analyses report exact, not asymptotic, p-values. (This choice does not affect any data interpretations.)

event set in the 3:4 condition, (*Mdn* = .125, 96% CI = [.00, .375]), *U* = 308.500, *p* < .001, *r* = .54. And likewise, the proportion of inner pad signals (which reflects use of the proximity-based mapping) was significantly greater for dyads in the 3:4 condition (*Mdn* = .875, 96% CI = [.625, 1.00]) than for dyads in the 4:3 condition (*Mdn* = .00, 96% CI = [.00, .25]), *U* = 308.500, *p* < .001, *r* = .54.

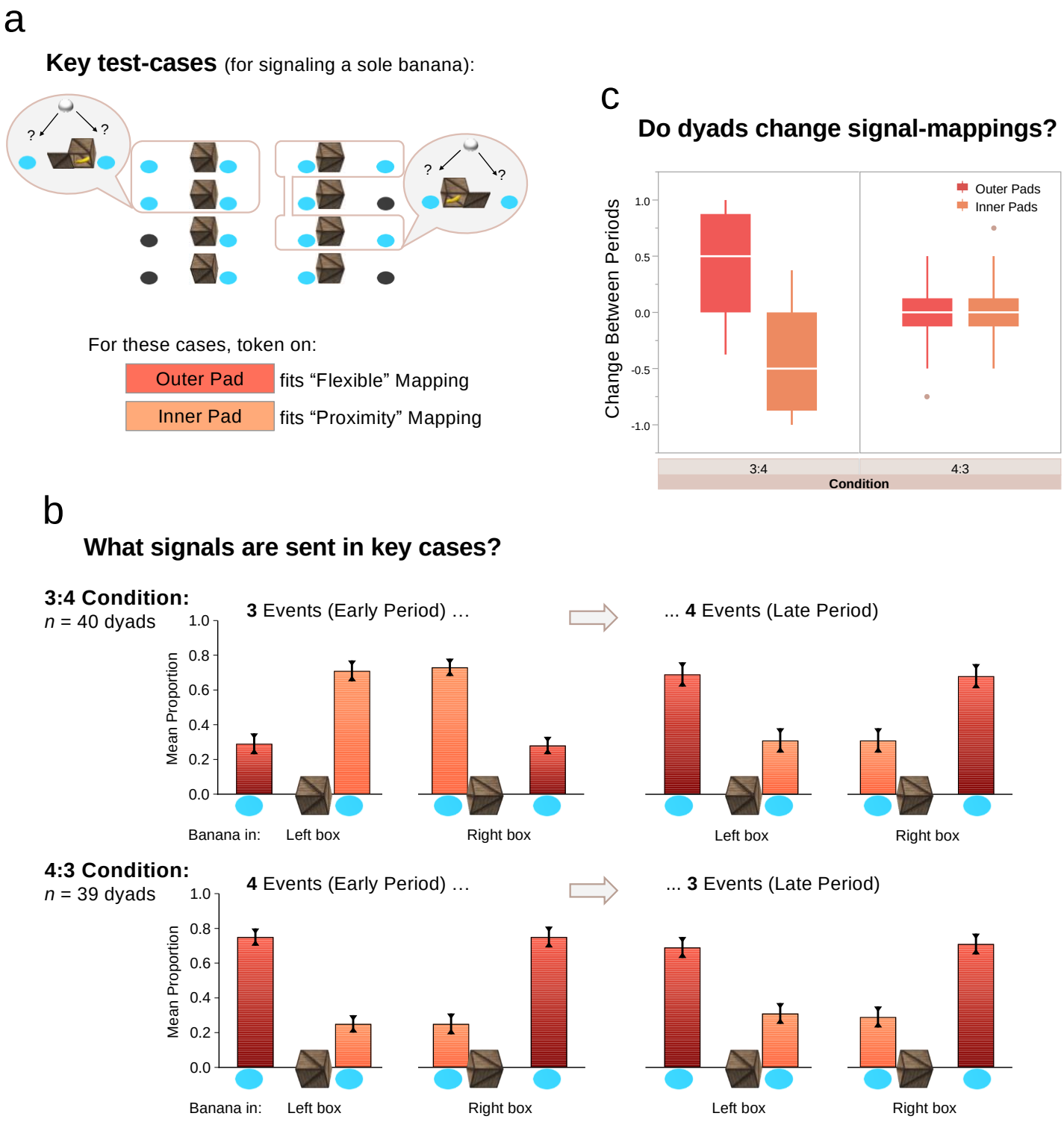

Fig. 6.

Signals sent by dyads to convey the location of a single banana in key test-cases. In both phases of the communicative task, there were (a) 4 cases testing between dyads' use of two

hypothesized mapping-types. For signaling the location of a sole banana in these particular contexts, dyads' choice of the Outer pad comports with the flexibly-oriented mapping; whereas dyads' choice of the Inner pad comports with the proximity-based mapping. Bar graphs (b) show the group-averaged proportions of signals that dyads transmitted in these key cases of the experiment, by condition and period. An image of the boxes-and-pads layout is superimposed on the x-axis, with bars projecting over pads to denote the associated proportions of signals sent for when the banana was in the left-box (left-half of each of the four graphs) or for when the banana was in the right box (right-half of each graph). Error bars represent ± 1 SE. Tukey outlier box plots (c) display how much individual dyads within each condition changed their use of Inner/Outer signals from the early to late period; that is, they show the distribution of pairwise differences, computed as the proportion of Inner [or Outer] pad signals from the late period minus the proportion of signals in kind from the early period. The white horizontal band in the boxes show the median, and whiskers extend to the lowest/highest datum within the distance of 1.5 * IQR (interquartile range) past the lower/upper quartile.

Next, to see whether changing the event-set across time also affected signal-meaning mappings, we compared signaling between the two periods within each condition. Consistent with switching from proximity-based to flexibly-oriented mappings, dyads in the 3:4 condition sharply decreased their use of inner pad signals from the 3-event early period (*Mdn* = .875, 96% CI = [.625, 1.00]) to the 4-event late period (*Mdn* = .00, 96% CI = [.00, .25]), $T = 20$, $p < .001$, $r = .57$. Correspondingly, 3:4 dyads' use of outer pad signals increased commensurately between these periods (3-event early period: *Mdn* = .125, 96% CI = [.00, .375]; 4-event late period: *Mdn* = 1.00, 96% CI = [.75, 1.00]), $T = 20$, $p < .001$, $r = .57$. Dyads in the 4:3 condition, however, did not change their mappings, with very high use of outer pad signals (flexibly-oriented mapping) persisting from the 4-event early period (*Mdn* = 1.00, 96% CI = [.75, 1.00]) into the 3-event late period (*Mdn* = .875, 96% CI = [.625, 1.00]), $T = 93$, $p = .284$, $r = .17$. Concomitantly, inner pad use remained low among these dyads (4-event early period: *Mdn* = .00, 96% CI = [.00, .25]; 3-event late period: *Mdn* = .125, 96% CI = [.00, .375]), $T = 93$, $p = .284$, $r = .17$.

Figure 6c shows the extent of these pairwise changes (or lack thereof), displaying the distributions of differences in signal use that individual dyads exhibited within each condition. In the 3:4 condition, the medians for dyads' pairwise change in proportions are +.50 for outer pad signals and -.50 for inner pad signals. In the 4:3 condition, the medians for dyads' pairwise change in proportions for either signals are 0.

2.3.2. Changing the set of signals.

We conducted additional analyses to assess the effect of changing the signal-set, as well as to confirm that dyads' full signals cohered as part of the system-level mappings stipulated.

Whereas individual signals might be predicted by more than one mapping-type, the collection of a dyad's signals within a block constituted a distinctive full mapping (specifying all possible phase-events) for each set of signal options. Accordingly, the signals that each dyad generated for each pad-set were considered together, by block, and classified as a mapping-type. In doing so, we extended the proximity-based mapping hypothesized in Fig. 5b (which covered signals only associated with the 3-event phase) to also cover signaling for the additional double-banana contingency in the 4-event phase. If dyads' construction of the proximity-based mapping is consistent with preemptive principles, then dyads can only signal "both" by choosing any blue pad that is *not* one of the two inner pads, which are used for "left" and "right" meanings. Alternatively, if the proximity-based strategy is motivated by an independent mapping approach, then we would expect the signal for "both" to be the same as "left" or "right." We dubbed the latter approach as "undifferentiated" and included this as a possible mapping-type response for the 4-event phase.

We omitted the fourth pad set (with only inner pads available) from these analyses, as predicted signaling is identical for all mapping-types. However, we note in passing that senders could have increased the specificity of their signals in this context (i.e., from "at least this box" → "both boxes" or "only this box") by implementing time delays in submitting signals (i.e., modulating the timing or duration of their communicative 'action' to differentiate meanings; see De Ruiter et al., 2010; Vesper et al., 2017). And indeed, at least a couple participants spontaneously reported, post-experiment, having enriched their signals this way.

Therefore, for each predictively-differential pad-set, per block, we coded dyad's signals as perfectly conforming (without deviation) to either a flexibly-oriented, proximity-based, or undifferentiated mapping-type; or if neither of these, then as "other." (For clarity, the definitions of these mapping-types are encapsulated in Figure 7.) Then, for each dyad, we computed the proportion of such full mappings in which each mapping-type was used, by event-phase.

Figure 7 shows the group-averaged proportions of mapping-types that dyads used within each period and condition. Flexibly-oriented and proximity-based mapping-types composed the vast majority, together totaling between .78 and .99 of dyads' average mapping proportions for any period within a condition. Thus, dyads strongly favored relational mapping approaches, in which signal-meaning pairs were contrastive, over mappings with independently-best but undifferentiated pairs, which were used rarely (with proportions of .04 and .09 for the undifferentiated mapping-type in the 4-event phases).

Binomial sign tests, comparing flexibly-oriented and proximity-based use, were conducted to ascertain the dominant mapping-type in the different periods. In accord with mapping-type definition, the flexibly-oriented mapping involves a reordering of mappings for

each pad-set; so substantial and dominant levels of this mapping-type would indicate that dyads' signal-meaning mappings change as the set of signaling options changes.

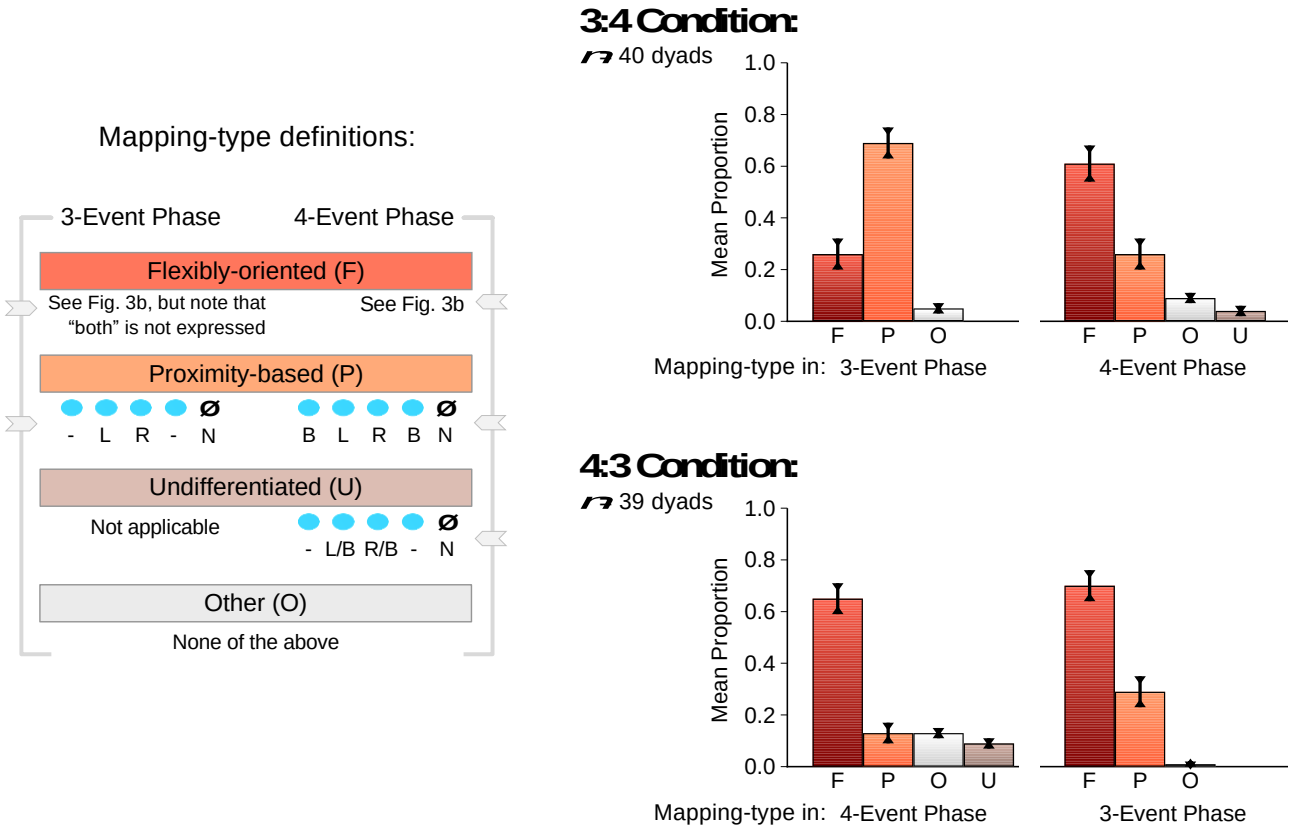

Fig. 7.

Mapping-types used by dyads for signaling**.** On the left, definitions are given under the horizontal bar for each mapping-type, and with respect to the 3-event and 4-event phases separately. Circles represent the four pads (from left-to-right sender's orientation) in the communicative task, along with the possibility of not signaling with any pad (∅). **L**etters beneath pads indicate which box(es) would be signaled by use of that pad as containing a target: *L* = left [box], *R* = right, *B* = both, *N* = none; a dash (-) indicates that no meaning is assigned to the pad and a slash (/) indicates two meanings associated with the same pad. Graphs show the group-averaged proportions with which dyads used the various mapping-types, by condition and

event-phase**.** Letter abbreviations under bars (and color-coding) correspond to mapping-type labels delineated at left side. Error bars represent ± 1 SE.

Indeed, in the 4-event early period, 4:3 dyads' proportions of using the flexibly-oriented mapping-type (*Mdn* = .67, 96% CI = [.50, 1.00]) dominated the proximity-based mapping-type (*Mdn* = .00, 96% CI = [.00, .00]), $S = 8$, $p < .001$; with 30 dyads using flexibly-oriented mappings more, 8 dyads using proximity-based mappings more, and 1 dyad using the two types equally. This pattern also carried over into 4:3 dyads' signaling in the 3-event late period (for the flexibly-oriented proportion, *Mdn* = .83, 96% CI = [.67, 1.00]; for the proximity-based proportion, *Mdn* = .17, 96% CI = [.00, .33]), $S = 8$, $p = .001$; 28 dyads preferred flexibly-oriented mappings, 8 dyads preferred proximity-based mappings, and 3 dyads used the two types equally.

Consistent with expectations, however, proximity-based mappings dominated over flexibly-oriented ones for 3:4 dyads in the 3-event early period (for proximity-based, *Mdn* = .83, 96% CI = [.50, 1.00]; for flexibly-oriented, *Mdn* = .08, 96% CI = [.00, .33]), $S = 7$, $p = .001$; among dyads, 28 preferred proximity-based mappings, 7 preferred flexibly-oriented mappings, and 5 showed no preference in use. The dominance of the proximity-based mapping-type disappeared, however, in the 4-event late period, which saw a switch towards 3:4 dyads' greater use of flexibly-oriented mappings (*Mdn* = .83, 96% CI = [.50, 1.00]; vs. proximity-based: *Mdn* = .00, 96% CI = [.00, .17]), $S = 12$, $p = .017$; only 12 dyads now used proximity-based mappings more, whereas 28 dyads used flexibly-oriented mappings more.

2.3.3. Instantaneity of system conventions.

Are dyads' construction of systems of conventions *instantaneous?* And similarly, does re-mapping from one system to another happen instantaneously? For both questions, the high levels of use of the flexibly-oriented system-mapping in relevant periods implies that some individuals must be doing so. And this observation alone—that people are indeed capable of instantly forming and modifying systems of conventions—is an important factor to account for. However, we can and have more closely examined, post-hoc, evidence for instantaneity in participants' signaling, as reported below.

Because a full mapping, or set of key mappings, is required to unequivocally confirm that a dyad is adopting a particular system mapping in this paradigm, it is not straightforward to conclude from a single trial alone whether a system-convention is being instantly adopted or switched. Despite this limitation, we can nonetheless look at some targeted measures that provide estimates for how common such instantaneity was. For these purposes, we focused on the flexibly-oriented mapping, which by its nature provides the clearest evidence in this task of system construction, and on the periods in which the flexibly-oriented mapping predominated in participants' signaling.

Firstly, in cases where dyads implemented a flexibly-oriented mapping across *all* trials within key (first) blocks, we can be assured that the mapping was effectively instantiated instantly. The frequency of such 'perfect' mapping instances thus provides a lower-bound estimate of instantaneity.

However, while the 'perfect' instantiation of a system-level convention implies instantaneity, such flawlessness underestimates the number of dyads who may have instantiated a system on the majority of trials but who may have deviated from that convention on occasion. We can therefore pinpoint the earliest trials in which the flexibly-oriented and proximity-based

signals differ and examine participants' responses on these 'first distinguishing' trials. The caveat that no one individual trial is definitive applies here, as the flexibly-oriented or proximity-based response may overlap on a given occasion with that from the "undifferentiated" or "other" mapping-types. (Although it is equally important to note that "undifferentiated" and "other" mapping-types were observed overall at minority frequencies per Fig. 7, and might alternatively ensue from switching between flexibly-oriented and proximity-based approaches). Accordingly, across the first three distinguishing trials, the frequencies of flexibly-oriented signals provide a reasonable upper-bound gauge of 'instantaneity.'

Next, we outline the application and findings from these lower-bound and upper-bound measures for instantaneous adoption and instantaneous switching, in turn. As with the signal-set analyses associated with Fig. 7, we will continue throughout to disregard responses from the fourth-pad set, which are not predictive of mapping-type, for these purposes.

2.3.3.1. Instantaneous adoption.

In accordance with the 'lower-bound' rationale sketched above, we identified the number of dyads that generated a 'perfect' flexibly-oriented mapping across all three predictively-differential pad-set configurations in the first block of the first periods in which the flexibly-oriented mapping predominated. These blocks corresponded to the absolute first block [in the early period] of the 4:3 condition and the first block of the *late period* of the 3:4 condition. In the 3:4 condition, 40% of dyads (16) perfectly, and thus instantly, instantiated the flexibly-oriented system convention. In the 4:3 condition, 35.9% of dyads (14) perfectly and instantly generated the flexibly-oriented mapping.

Applying our ‘upper-bound’ criterion, we also computed the percentage of signals consistent with the flexibly-oriented mapping on the first three distinguishing trials occurring within these same first blocks. ‘Distinguishing trials’ were those in which flexibly-oriented and proximity-based mappings differed, corresponding to the key cases identified earlier (see Fig. 6a) for signaling the location of a single banana *and* also those cases when there were two bananas present (since the relevant blocks are both in the 4-event phase). The number of sent signals for each mapping-type on the first-three distinguishing trials are depicted in Figure 8. Suggestive of instantaneous system generation, the percentage of signals consistent with the flexibly-oriented mapping was 80% (32), 62.5% (25) and 70% (28) for the 3:4 condition, and 82.1% (32), 71.8% (28) and 61.5% (24) for the 4:3 condition.

Signals sent on the first trials in which 'flexible' and 'proximity' mappings differ.
(4-event phases only)

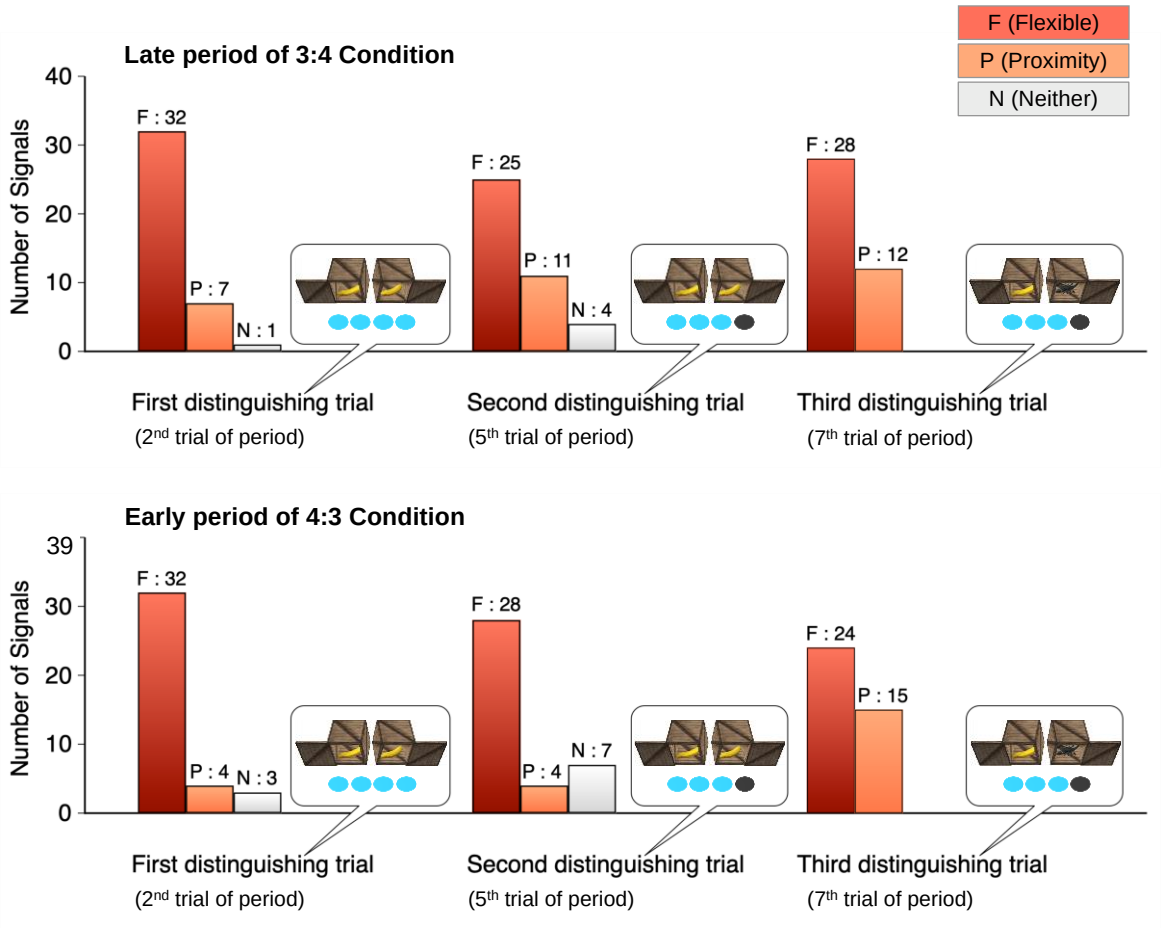

Fig. 8.

Number of signals that comported with the flexibly-oriented (F; 'flexible') or proximity-based (P; 'proximity) mapping, or neither (N), on the first three occurrences in which such mappings were distinguishable from each other. Such 'distinguishing' trials were examined within the 4-event phase only of both experimental conditions, which corresponded to the late period of the 3:4 condition (top graph) and the early period of the 4:3 condition (bottom graph). Callouts depicted with each instance indicate the scene-event (e.g., two bananas present) and signaling set (schematic arrangement of blue and black pads, as viewed from left-to-right sender's orientation) that was observed on that trial. Note that callouts are identical between conditions here as, per design, the same ordering of trial-types was used for the 4-event phase.

2.3.3.2. Instantaneous switching.

As reported earlier, we observed that dyads completely modified a system of mappings when the meaning-set changed in the 3:4 condition and as the signal-set changed in the 4-event phases. How often does such switching from one system-mapping to another occur instantaneously?

To evaluate 'switching' when the meaning-set changed, we first identified those dyads who generated a 'perfect' *proximity-based* mapping in the last block prior to the change in meanings in the 3:4 condition (i.e., across all three predictively-differential pad-set configurations in the second block of the first period in the 3:4 condition). Twenty-three of the 40 dyads met this criterion. From among this subset of 23 dyads, we then identified the proportion that subsequently switched to a 'perfect' flexibly-oriented mapping [across all three predictively-differential pad-set configurations] in the first block after the meaning-set change. Eight of the 23 dyads' signals conformed to a perfect flexibly-oriented mapping, resulting in 34.8% of dyads with a perfect proximity-based system instantly switching to a perfect flexibly-oriented system.

As before, we also looked at responses on the first distinguishing trials for evidence of less flawless, but nonetheless suggestive of instantaneous, system-switching. These are the same distinguishing trials described above and depicted in Figure 8, but this time, to evaluate instantaneous switching associated with changes in the meaning-set, we only looked at responses sent by dyads whose dominant mapping in the prior block was the proximity-based one. That is, among dyads in the 3:4 condition who generated the proximity-based mapping for at least two of the three pad-set configurations in the last block of the early period, we examined the signals they subsequently sent on the first distinguishing trials of the first block of the late period. Table 1 shows the breakdown of their signals by mapping-type. Across each of the three distinguishing

trials, 76.7%, 53.3% and 63.3% of the dyad subset (n = 30 dyads) switched to a signal consistent with the flexibly-oriented mapping.[17]

For switching associated with changes in the signal-set, notice that the flexibly-oriented mapping itself entails a re-mapping as the signal set changes. Therefore, dyads instantiating a perfect flexibly-oriented mapping in the first block of the 4:3 condition are also demonstrating instant switching in this regard. This corresponded to 35.9% of dyads in this condition (as reported earlier for the lower-bound estimate on instantaneous adoption). Similarly, as the first distinguishing trials involved different pad-set configurations, the percentage of 'flexibly'-consistent signals on these trials is also suggestive of instant switching as the signal-set changes: with 71.8% and 61.5% of the 4:3-condition dyads (as reported above) sending signals consistent with the flexibly-oriented mapping on the second and third distinguishing trials, respectively.

In sum, figures of *perfect* system-level adoption at 35.9 and 40%, and estimates of instant, flexibly-consistent signals in the range of 61.5 to 80%, suggest a substantial degree of instantaneity in the generation of system-level conventions. Likewise, the figures of perfect switching at 34.8 and 35.9%, and the estimates of flexibly-consistent signal-switching at 53.3 to 76.7%, also imply a substantial degree of instantaneity in switching from one system of conventions to another.

[17] In evaluating instantaneous switching associated with changes in the meaning-set, a potential concern is that 'the switch' may happen because dyads alternate Sender roles across the two relevant blocks (so maybe we are just capturing differences in the preferences of individual Senders, rather than a re-mapping process per se). However, when one also evaluates these measures in relation to the Sender-first member of each dyad (i.e., changes in signals by the Sender from the first block, in the early 3:4 period, to the same Sender in the third block, in the late 3:4 period), the overall patterns remain the same in support of instantaneous switching. Namely, 8 out of 24 senders (33.3%) with a perfect proximity-based mapping switched to a perfect flexibly-oriented mapping; and for the corresponding analysis regarding first distinguishing trials, the percentages of senders switching to a flexibly-consistent signal was 69.2%, 53.8% and 61.5%.

Table 1

Number and percentage of signals for the 'first distinguishing trials' for gauging instantaneity of system conventions.

	First distinguishing trial						**Second distinguishing trial**						**Third distinguishing trial**					
	Flexible		Proximity		Neither		Flexible		Proximity		Neither		Flexible		Proximity		Neither	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Instantaneous adoption, 3:4 condition (dyad *n* = 40)	32	80.0	7	17.5	1	2.5	25	62.5	11	27.5	4	10.0	28	70.0	12	30.0	0	0
Instantaneous adoption, 4:3 condition (dyad *n* = 39)	32	82.1	4	10.3	3	7.7	28	71.8	4	10.3	7	17.9	24	61.5	15	38.5	0	0
Instantaneous switch, with change in meaning-set (dyad *n* = 30)	23	76.7	6	20	1	3.3	16	53.3	10	33.3	4	13.3	19	63.3	11	36.7	0	0
Instantaneous switch, with changes in signal-set (dyad *n* = 39)	32	82.1	4	10.3	3	7.7	28	71.8	4	10.3	7	17.9	24	61.5	15	38.5	0	0

Note. (Blue-)shaded columns correspond to the percentage of signals consistent with the flexibly-oriented system-mapping, which is the key mapping-type for strongly evidencing instantaneity of system conventions in this paradigm. Numbers for "Instantaneous adoption" are also graphed in Figure 8, alongside additional information on the composition of the first distinguishing trials. Figures for "Instantaneous switch, with changes in signal-set" are necessarily synonymous with those for "Instantaneous adoption, 4:3 condition" (see main text for explanation); they are reduplicated here for clarity.

2.4. Discussion

In a non-linguistic signaling game, participants created novel systems of signal-meaning mappings, in which the meaning of each signal contrasted with the meanings of other available, but unchosen, signals. Crucially, the system-level nature of such mappings was underscored by participants' systematic, and often instantaneous, re-mapping of signal-meaning pairs in instances where the set of signals, or set of meanings, changed. Specifically, when the set of signals changed (within the 4-events phase), mappings were spontaneously reordered in the moment. And when the set of meanings differed between conditions (in the early periods), mappings changed, too: the interpretation of any one signal (e.g., for designating the location of a single banana) depended on the set of other possible meanings in the current situation. Furthermore, when the set of meanings changed (across periods), participants remapped entirely when their existing mapping-system was inadequate for the new environment. Notice that from a joint action and reasoning perspective, this theoretically implies not only that the sender and receiver both recognize that remapping would be advantageous, but also that the parties take this to be common ground, so that they can reliably adopt the new mapping, expecting that the other has done the same.

This flexible modification of systems of signal-meaning mappings is ubiquitous in communication and natural language. The driving scenario in the Introduction provides one of many examples for human communication on the roadways, using a simple visual channel of signals (though, of course, other modes of signaling are available and expand these possibilities). In language, some adjectives of degree – such as those pertaining to size, depth, costliness, thinness, height, and speed – routinely show re-mappings, when applied to different sets of objects. What a "large" coffee refers to exactly at the local cafe depends on the sizes available.

More broadly, the phenomenon of scalar implicature may reflect counterfactual reasoning. Scalar expressions typically function contrastively (*some* vs. *all, warm* vs. *hot, like* vs. *love*), despite semantic compatibility: the use of a less specific expression is often interpreted pragmatically to mean that the narrower, available alternative does not apply (or it would have been used).

Preferred system mappings in natural language can be linguistically diverse, as may be produced from the variable coding of spatial relationships (Majid et al., 2004). Different, conflicting systems for the same domain may even be conventionalized within the same language: e.g., the mapping of space-time metaphors in English has produced both 'ego-moving' and 'time-moving' systems for talking about events in time, sometimes yielding different interpretations for the same utterance (e.g., 'the meeting is moved *forward*') (Gentner & Imai, 1992; Boroditsky, 2000).

The flexibility with which people are able to switch to new systems of conventions when required by the demands of the communicative situation raises the question of how far even deeply entrenched communicative conventions may potentially be over-ridden spontaneously, when required by demands of the communicative situation. Indeed, simple examples of this are familiar, even for very established conventions. For example, in a situation in which outright rejection might seem impolite, a delayed and slow nod of the head may be interpreted as a 'No'—contrastively with a typical nod, which would be interpreted affirmatively. And this remapping, too, is provisional: for a person visibly recovering from a neck strain, no such remapping would seem appropriate—the less-than-vigorous nod would have an alternative, medical explanation.

This study is the first demonstration, to our knowledge, of participants' *de novo* creation of

systems of novel conventions (without using languagelike stimuli, familiar gestures or signs, or pre-existing communicative systems), which are evidenced by their flexible dependence on counterfactual reasoning about alternative signals and meanings. A partial reflection of this phenomenon can be seen in cases involving minimal contrast between alternate signals. For example, Bergen, Goodman and Levy (2012) found that people in a signaling game readily coordinated on mapping a set of two possible signals (triangle, 'alien' symbol) onto two meanings (cube, pyramid) by reasoning that the alien symbol should be paired with the cube because a more specific, iconic relationship would obtain between the triangle and pyramid.

Our study also found preliminary evidence of hysteresis for system-mappings in the 4:3 condition, in which dyads generated early on a 'good enough' mapping from which there was no reason to later change, when the 'world' became simpler to accommodate. Thus, people do not necessarily switch to the most 'natural' available mapping in the situation (if that situation were considered without precedent), but continue to use their current mapping if it is adequate to the demands of the moment. This is reminiscent of patterns shown in the experimental pragmatics literature on conversational partners (Brennan & H. H. Clark, 1996): Once dyads establish partner-specific lexical conventions for referring to objects, they are more likely to continue using such 'conceptual pacts' (if adequate), even when simpler lexical choices for disambiguating references become possible. Although in our study, dyads aimed to coordinate their actions under the impression of 'community'-wide exchanges (rather than through known partner-specific pacts), dyads [in the 4:3 condition] transferred the early system-mapping convention despite the possibility of using a simpler, and *ahistorically* preferable, system-mapping for the later period. (In contrast, dyads in the 3:4 condition readily reorganized their system-mappings in the reverse case under which their mappings became functionally

inadequate for the later environment.) These results point, then, to a possible role for precedence and common history in influencing when a linguistic convention, or system of conventions, changes or remains stable.

The notion of 'systematic' mappings is notably distinct from the notion of systematicity that is commonly used in the language literature and that has been demonstrated in studies of experimental semiotics and iterated learning (Kirby et al., 2008; Theisen et al., 2009). Those findings showed how participants produced regularities *within* signs or signals (regarding a component or feature of its form) that were predictive of aspects of meanings. Conversely, our findings show how systematicity applies *across* signal-meaning pairs *in toto*, seen here in the mapping of atomic signals to the world. People in our study created signal-meaning mappings that were globally cohesive, and interdependently adapted to changes in communicative constraints.

More broadly, our study is consistent with social cognition approaches to pragmatics (Bergen et al., 2012; H. H. Clark, 1996; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Grice, 1975; Sperber & Wilson, 1986), and highlights the likely operation of an 'interactional intelligence' (Levinson, 2006) that is foundational to communicative conventions. Without using pre-established forms of communication, interacting partners in a coordination task used processes of social and *joint inference* (see also Misyak et al., 2016)—manifesting most strikingly in the ability to instantaneously create and modify novel systems of conventions.

3. General Discussion

Human communication involves the flexible construction of links between signals and meanings, to achieve the communicative goals of the moment—the creation of what we term "instantaneous" conventions, which may either be variations on pre-existing conventions (as in signals in driving, or in natural language) or entirely new signals invented for the occasion. Moreover, interpretation of each signal contrasts with the interpretation of other possible signals, so that the conventions operate as a system. In this paper, we explore and address theoretical challenges in explaining how instantaneous conventions can work. These do not appear easy to explain using some accounts of pragmatic inference, which do not have mechanisms for choosing between possible conventions; but they can be explained using the theory of virtual bargaining (Misyak et al., 2014), according to which people choose a convention by reasoning about which convention would be agreed, if they could communicate. Moreover, we have shown empirically that people can, indeed, generate instantaneous systems of communicative conventions which are shaped by the task demands of the moment.

This work has potentially interesting implications for the three related research traditions mentioned earlier: experimental pragmatics, experimental semiotics and work on the cultural evolution of language. Regarding experimental pragmatics, we suspect that the reasoning that can allow people to coordinate on novel instantaneous conventions will also be operative in standard pragmatic inference, in which language is frequently used in highly flexible, and non-conventional, ways (e.g., E. V. Clark & H. H. Clark, 1979; H. H. Clark, 1996). Similarly, regarding experimental semiotics, the complex reasoning that can underpin a one-shot communication with a novel non-linguistic signal is likely to shape communication where successive interaction with iterative feedback is possible, and to help understand processes of adjustment and repair by which partners rapidly converge on a usable convention (Galantucci,

2005; Healey et al, 2007). Finally, we suggest that the active, creative ways in which systems of conventions can be invented in the moment is likely to guide the cultural evolution of linguistic systems. Rather than seeing linguistic conventions as shaped purely by backward-looking processes of selection (e.g., through imitating or reinforcing patterns that prove successful and eliminating those that do not [Skyrms 2010; Nowak, Krakauer & Kingdom, 1999]), it may be more appropriate to see the creation of language as a process of forward-looking, active reasoning. This reasoning is, though, focused purely on the communicative demands of the moment—the gradual entrenchment of increasingly stable conventions arises purely as a side-effect (Christiansen & Chater, 2022; Tomasello, 2010).

Building on Clark (1960), we have argued that the flexible reasoning involved in creating novel systems of communicative conventions is grounded in a process of virtual bargaining: a process by which sender and receiver simulate a negotiation concerning which communicative convention to use on the basis of their common ground. We stressed above that the chosen convention can depend on an indefinitely large range of factors, including the nature of the environment, the set of signals available and potential meanings to be communicated, potential outcomes for each player, their preferences, bargaining power, history of past interactions and so on. Can such an approach be modelled formally? We suggest that any such modeling will be limited in scope. The process of establishing what is in common ground, the problem of creatively inventing possible conventions, and the general process of negotiation between people, are all problems that appear to depend on human intelligence quite broadly, and hence lie outside the scope of a cognitive model. Nonetheless, within restricted contexts (such as in experiments in which signals, layouts, and outcome payoffs can be specified), building a formal

model seems tractable, particularly in the light of recent related work (e.g., Bundy, Philalithis & Li, 2021; Stacy et al., 2021; Wang et al., 2021).

Finally, we suggest that virtual bargaining may provide a possible foundation for conversational pragmatics, in the Gricean tradition. Social interaction in general, and communication in particular, is often viewed as a process of creative and continually changing negotiation (e.g., Garfinkel, 1967); and much, and perhaps most, such negotiation is implicit (e.g., Polanyi, 1966). The virtual bargaining approach to communication takes this viewpoint as the starting point for understanding the human ability to communicate using novel signals and in new situations. Our remarkable ability to create and use "instantaneous" systems of conventions in one-off communicative interactions may, from this point of view, be the starting point from which the stable conventions underlying human language ultimately emerge.

Acknowledgments

We thank Ty Hayes for programming support.

Funding sources

JM and NC were supported by UK- EPSRC grant EP/N012380/1 and ERC grant 295917-RATIONALITY. NC was also partially supported by the ESRC Network for Integrated Behavioural Science grant ES/K002201/1, Leverhulme Trust grant RP2012-V-022, and Research Councils UK Grant EP/K039830/1.

Supplementary material

Raw data, processed data files, and codebooks for the experiment are publicly available at https://osf.io/t58ys (OSF repository).

References

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5-6), 509-559.

Bacharach, M. (2006). Beyond individual choice: Teams and frames in game theory. (N. Gold & R. Sugden, Eds.). Princeton, NJ: Princeton University Press.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 120-125).

Berkowitz, L., & Walster, E. (Eds.). (1976). *Advances in experimental social psychology (Vol. 9), Equity theory: Toward a general theory of social interaction*. New York, NY: Academic Press.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition, 75*, 1-28.

Brennan, S., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482-1493.

Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews, 2*(3), 177-226.

Brochhagen, T. (2020). Signalling under uncertainty: Interpretative alignment without a common prior. *The British Journal for the Philosophy of Science, 71*(2), 471-496.

Bundy, A., Philalithis, E., & Li, X. (2021). Modelling Virtual Bargaining using Logical Representation Change. In. S. Muggleton, S. & N. Chater, N. (Eds). *Human-like machine intelligence* (pp. 68-89). Oxford, UK: Oxford University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711-733.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*(7), 1131-1157.

Chater, N. & Misyak, J. (2021). Spontaneous communicative conventions through virtual bargaining. In. S. Muggleton, S. & N. Chater, N. (Eds). *Human-like machine intelligence* (pp. 52-67). Oxford, UK: Oxford University Press.

Chater, N., Misyak, J. B., Melkonyan, T., & Zeitoun, H. (2016). Virtual bargaining: Building the foundations for a theory of social interaction. In: J. Kiverstein (Ed.), *Handbook of the philosophy of social mind* (pp. 418- 430). New York & Oxford: Routledge.

Chater, N., Misyak, J., Watson, D., Griffiths, N., & Mouzakitis, A. (2018). Negotiating the traffic: Can cognitive science help make autonomous vehicles a reality? *Trends in Cognitive Sciences*, *22*(2), 93-95.

Chater, N., Zeitoun, H., & Melkonyan, T. (2019). The social character of moral reasoning (Commentary on Joshua May, Regard for Reason in the Moral Mind). *Behavioral and Brain Sciences, 42*, E149. doi:10.1017/S0140525X18002583.

Chater, N., Zeitoun, H., & Melkonyan, T. (in press). The paradox of social interaction: Shared intentionality, we-reasoning and virtual bargaining. *Psychological Review*.

Christiansen, M., & Chater, N. (2008). Language as shaped by the brain (with commentaries and reply). *Behavioral and Brain Sciences*, *31*, 489-558.

Christiansen, M. & Chater, N. (2016). *Creating Language*. Cambridge, MA: MIT Press.

Christiansen, M. H. & Chater, N. (2022). *The language game*. London, UK: Bantam Books/New York, NY: Basic Books.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317-35.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language, 17*(2), 417-431.

Clark, E. V., & Clark, H. H. (1979). When nouns surface as verbs. *Language, 55*, 767-811.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1-39.

Colman, A. M. & Gold, N. (2018). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review, 25,* 1770-1783.

Croft, W. (2000). *Explaining Language Change. An Evolutionary Approach*. Harlow: Pearson Education.

Culicover, P. W. (1999). *Syntactic nuts: Hard cases, syntactic theory, and language acquisition*. Oxford, UK: Oxford University Press.

Degen, J., & Tanenhaus, M. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*, 172–201.

De Ruiter, J. P., Noordzij, M. L., Newman-Norlund, S., Newman-Norlund, R., Hagoort, P., Levinson, S. C., & Toni, I. (2010). Exploring the cognitive infrastructure of communication. *Interaction Studies*, *11*, 51–77.

Feldman, J., & Choi, L. S. (2022). Meaning and reference from a probabilistic point of view. *Cognition*, *223*, 105058.

Fodor, J. A., Garrett, M. F., Walker, E. C., & Parkes, C. H. (1980). Against definitions. *Cognition, 8*(3), 263-367.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games.

Science, *336*, 998–998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*(1), 80–96.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*, 737–767.

Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental semiotics. *Language and Linguistics Compass*, *6*, 477–493.

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialog: A study in conceptual and semantic coordination. *Cognition*, *27*(2), 181–218.

Garrod, S., & Doherty, G. (1994). Conversation, coordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3), 181–215.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science, 31,* 961-987.

Gentner, D. & Ima, M. (1992). Is the future always ahead? Evidence for system-mappings in understanding space-time metaphors. *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 510-515).

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389-407.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251-279.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–-829.

Goodman, N., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*, 377–88.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). New York: Academic Press.

Hanna, J.E., Tanenhaus, M.K., & Trueswell, J.C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory & Language, 49*, 43–61.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, *74*(1), 88-95.

Hart, H. L. A. (1961). *The Concept of Law*. Oxford: Clarendon Press.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science, 44*(6), e12845.

Hawkins, R. D., Franke, M., Frank, M. C., Smith, K., Griffiths, T. L., & Goodman, N. D. (2021). From partners to populations: A hierarchical Bayesian account of coordination and convention. arXiv preprint arXiv:2104.05857.

Hawkins, R. X., Franke, M., Smith, K., & Goodman, N. D. (2018). Emerging abstractions: Lexical conventions are shaped by communicative context. In T. M. Rogers, M. Rau, J. Zhu & C. Kalish (Eds.) *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 463-468). Red Hook, NY: Curran Associations.

Healey, P. G. T., Garrod, S., Fay, N., Lee, J. & Oberlander, J. (2002). Interactional context in graphical communication. In W.D. Gray and C.D. Schunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 441-446). Mahwah, NJ: Erlbaum.

Healey, P. G. T., Swoboda, N., Umata I. and King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science, 31*, 285-309.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Hoffman, E., & Spitzer, M. L. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *The Journal of Legal Studies*, *14*(2), 259-297.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge, UK: Cambridge University Press.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground?. *Cognition*, *59*(1), 91-117.

Hu, J., Levy, R., & Zaslavsky, N. (2021). Scalable pragmatic communication via self-supervision. arXiv preprint arXiv:2108.05799.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45-52.

Keysar, B., Bar, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11,* 32-38.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105,* 10681-10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141*, 87-102.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

Lascarides, A., & Copestake, A. (1998). Pragmatics and word meaning. *Journal of Linguistics*,

34(2), 387-414.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. (2018). The Cognitive Mechanisms of Contractualist Moral Decision-Making. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. p. 683.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT press.

Levinson, S. C. (2006). On the human "interaction engine". In N. J. Enfield, & S. C. Levinson (Eds.), Roots of human sociality: Culture, cognition and interaction (pp. 39-69). Oxford: Berg.

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

MacWhinney, B. (2014). Item-based patterns in early syntactic development. In *Constructions collocations patterns.* In T. Herbst, H.J. Schmid & S. Faulhabe (Eds.) (pp. 33-70). Berlin: De Gruyter Mouton.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*, 108–114.

Melkonyan, T., Zeitoun, H., & Chater, N. (2018). Collusion in Bertrand versus Cournot competition: A virtual bargaining approach. *Management Science*, *64*(12), 5599-5609.

Misyak, J. B. & Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369,* 20130487. doi: 10.1098/rstb.2013.0487

[dataset] Misyak, J., & Chater, N. (2019). Data for “Instantaneous Systems of Communicative Conventions” - Misyak & Chater. OSF. osf.io/t58ys

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: Virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences, 18,* 512-519.

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science, 27,* 1550-1561.

Morin, O. (2018). Spontaneous emergence of legibility in writing systems: The case of orientation anisotropy. *Cognitive Science, 42*(2), 664-677.

Noveck, I., & Reboul, A. (2008). Experimental pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences, 12*, 425–431.

Noveck, I. A., & Sperber, D. (Eds.). (2004). *Experimental pragmatics*. Basingstoke: Palgrave Macmillan.

Nowak, M., Krakauer, D., & Kingdom, U. (1999). The evolution of language. *Proceedings of the National Academy of Sciences, 96*(July), 8028–8033.

Peloquin, B. N., Goodman, N. D., & Frank, M. C. (2020). The interactions of rational, pragmatic agents lead to efficient language structure and use. *Topics in Cognitive Science*, *12*(1), 433-445.

Polanyi, M. (1966). *The tacit dimension*. New York, NY: Doubleday.

Roberts, G. & Galantucci, B. (2012). The emergence of duality of patterning: Insights from the

laboratory. *Language and Cognition, 4,* 297-318.

Schelling, T. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford, UK: Oxford University Press.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226-233.

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

Sperber, D., & Hirschfeld, L. A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences*, *8*(1), 40–46.

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., Kleiman-Weiner, M., & Gao, T. (2021). Modeling communication to coordinate perspectives in cooperation. *arXiv preprint, arXiv:2106.02164*.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469-488.

Steels, L., & Loetzsch, M. (2012). The grounded naming game. In L. Steels (Ed.), *Experiments in cultural language evolution* (pp. 41–59). Amsterdam: John Benjamins.

Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations, 6,* 165-181.

Theisen, C. A., Oberlander, J., & Kirby, S. (2009). Systematicity and arbitrariness in novel communication systems. In N. Taatgen & H. van Rijn (Eds.) *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1971-1976).

Tomasello, M. (2003). *Constructing a language. A usage-based theory of language acquisition*. Cambridge, MA : Harvard University Press.

Tomasello, M. (2010). *Origins of human communication*. Cambridge, MA: MIT press.

van den Bosch, A. & Daelemans W. (2013). Implicit schemata and categories in memory-based language processing. *Language and Speech*, *56*(3):309-28.

Vesper, C., Schmitz, L., & Knoblich, G. (2017). Modulating action duration to establish nonconventional communication. *Journal of Experimental Psychology: General*, *146*, 1722–1737.

Waismann, F. (1945). Verification. *Proceedings of the Aristotelian Society (Supplementary Volumes)*, *XIX*, 119-150.

Wang, R., Wu, S., Evans, J., Tenenbaum, J., Parkes, D., & Kleiman-Weiner, M. (2021). Too Many Cooks: Coordinating multi-agent collaboration through inverse planning. In S. Muggleton & N. Chater (Eds.), *Human-like Machine Intelligence* (pp. 152-170). Oxford, UK: Clarendon Press.

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15-30.

Wittgenstein, L. (1953). *Philosophical Investigations*, G.E.M. Anscombe and R. Rhees (eds.), G.E.M. Anscombe (trans.), Oxford: Blackwell.

Zipf, G. K. (1935). *The psycho-biology of language* (Vol. ix) Oxford, England: Houghton Mifflin.