Trends in Cognitive Sciences

CellPress REVIEWS

Review

The Emergence of Social Norms and Conventions

Robert X.D. Hawkins,[1] Noah D. Goodman,[1,2] and Robert L. Goldstone[3,4,*]

The utility of our actions frequently depends upon the beliefs and behavior of other agents. Thankfully, through experience, we learn norms and conventions that provide stable expectations for navigating our social world. Here, we review several distinct influences on their content and distribution. At the level of individuals locally interacting in dyads, success depends on rapidly adapting pre-existing norms to the local context. Hence, norms are shaped by complex cognitive processes involved in learning and social reasoning. At the population level, norms are influenced by intergenerational transmission and the structure of the social network. As human social connectivity continues to increase, understanding and predicting how these levels and time scales interact to produce new norms will be crucial for improving communities.

Highlights

Much of our social world is governed by norms, which can have life or death consequences for the people who hold them. The behavior and beliefs of one agent depend in more or less complex ways on the often unspoken expectations held about other agents.

Social norms depend on multilevel, interactive processes that include internal cognitive processes within an individual as well as constraints on the communicative channels that connect people.

Norms can be both the consequence and facilitator of social interactions.

Navigating a Web of Social Expectations

Even a casual observer of humanity will be struck by the similarity of behavior displayed by individuals within a community, and the surprising variation across different communities. Where people come from, or where they currently spend their time, influences the language they speak, the clothes they wear, the food they eat, the currency they spend, and countless other routine behaviors: whether they eat hamburgers with their fingers (as in the USA) or knife and fork (as in Norway); eat rice with chopsticks (Japan) or their right hand (Malaysia); dip French fries in ketchup (Canada) or mayonnaise (Belgium); arrive to dinners on time (Germany) or fashionably late (Brazil); and sit in the back (England) or front (Australia) of a taxi cab during a solo trip. This influence extends beyond the ubiquitous social interactions of everyday life to decisions with potentially life-altering consequences: whether to challenge an adversary in a duel [1], reciprocate gang-related violence [2], donate one's organs [3], drink or smoke cigarettes [4], wear a helmet while riding a bicycle [5], or even to report sexual harassment, have a child, or allow a clitoridectomy or circumcision to be performed on one's child.

While some correlated behaviors may simply be chalked up to shared habits (e.g., we all brush our teeth each morning), what distinguishes a broad class of norms is the way agents mentally represent them. Key distinctions have been proposed between several related constructs to organize this complexity, and the precise relationships among conventional, descriptive, and prescriptive norms remain under debate (Box 1). Yet norms of all stripes share a common foundation. In each case, the behavior and beliefs of one agent depend in more or less complex ways on the often unspoken expectations held about the behavior and beliefs of other agents. Where these expectations come from, how they are represented in individual minds, and how they are sustained or shift in different populations over different time scales, are core questions for cognitive science, with broad ramifications for an increasingly interconnected society.

[1]Department of Psychology, Stanford University, Stanford, CA, USA
[2]Department of Computer Science, Stanford University, Stanford, CA, USA
[3]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA
[4]Cognitive Science Program, Indiana University, Bloomington, IN, USA

*Correspondence: rgoldsto@indiana.edu (R.L. Goldstone).

Box 1. Distinctions between Different Kinds of Norms

The concept of a norm is not monolithic. Extensive work in philosophy, social psychology, and developmental psychology has sought to tease apart distinct varieties, with important differences in how they are represented and affect behavior [10,112–115].

Descriptive norms, such as fashions or customs, only require that people tend to conform to the behaviors prevalent within their communities and have knowledge about what is prevalent. These are unilateral: if I choose to eat with an unusual utensil, it doesn't materially affect another agent's utility and my success in eating doesn't depend on that agent's expectations.

Prescriptive norms (or moral norms), such as shared notions of justice or fairness, are stronger expectations about what people ought to do. They may therefore take on additional moral or injunctive force (violators experience guilt and can expect to be punished [116]) and may also be viewed as less subjective and more likely to apply beyond the bounds of one's own community [117].

Conventions, such as which particular arrangement of gesticulated fingers signifies disgruntled vexation, require a bidirectional coordination of expectations in interaction: the sender and receiver must each expect that the other shares their interpretation for the interaction to succeed.

In practice, however, this taxonomy is often blurred in interesting ways. For example, what begins as a descriptive norm or convention may take on prescriptive force: For example, driving on the nonconventional side of the road is not simply miscoordinating but also regarded as normatively reckless and irresponsible due to various externalities. And despite decades of pleas from linguists taking a purely descriptivist stance, many ordinary language users continue to treat, in principle, conventional grammatical choices with moral fervor. Thus, the precise boundaries to carve between different kinds of norms, and the relationships between them, remain an exciting open area.

We synthesize the literature on the emergence of norms and conventions by viewing them as both governing and being governed by local social interactions (Figure 1). Existing norms provide initial constraints on local social interactions that, in turn, can modify the norms. The novel expectations people form during interaction are shaped by complex cognitive processes within individuals as well as the interaction channels across them. As soon as the first filaments of a local norm connect the individuals, these expectations will shape all subsequent interactions among them and often generalize to interactions with other partners, leading to broader changes. In focusing on these functional cognitive foundations, we present a complementary perspective to other recent reviews focusing more specifically on emergent, population level phenomena [6–8].

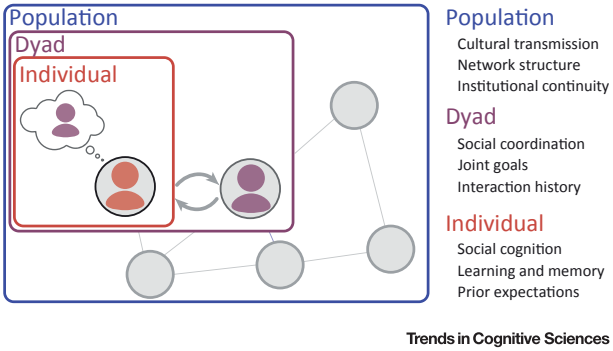


Figure 1. Illustration of Our Theoretical Perspective. Three levels of processes work together to form, perpetuate, and reshape norms in a community.

Why Do Norms Form? A Functional View

To successfully navigate the physical world, cognitive agents must form an understanding of how inanimate objects are expected to behave. The social world can pose an even more dizzying computational challenge. We take the philosopher David Lewis' analysis of conventions [9], the most influential in a long line of philosophical treatments [10–12], as a starting point for highlighting several key theoretical properties that not only illuminate how norms work but also why they may be useful for agents in the first place. In this framework, conventions are behavioral regularities that serve as stable but to some degree arbitrary solutions to repeated coordination problems. Using language to communicate is a paradigmatic example. Because we are not telepathic, we often find ourselves in the position of needing to use the sensory data we produce and perceive to refer to novel objects or ideas with novel partners (thus a repeated problem). Understanding each other requires both the speaker and listener to share roughly the same expectations about mappings between linguistic forms and meanings (thus a coordination problem). As attested by the great diversity of languages documented across the continents, or even across dialects of the same language (Figure 2), there are many possible solutions to this problem (arbitrariness), but working out a new mapping from scratch in every interaction would be extremely inefficient at best. Hence, once a particular solution is widely adopted, it is in everyone's best interest to keep using it (stability). We note that the conventionality of semantic mappings is uncontroversial, but despite increasing evidence for the conventionality of other aspects of language such as grammatical constructions [13], this view remains under debate [14].

While this analysis is specific to the subvariety of norms known as conventions, where it is already in each agent's best interest to coordinate, prescriptive norms have been understood as solutions to problems where coordination may not initially be in each agent's self-interest [10]. We adhere to such norms even when we are competing and have strong incentives to break them [15]. A striking example of this is the tacit agreement among British and German soldiers in World War I not to fire upon enemies when they were retrieving wounded or dead comrades, or at times, when they were simply resting, exercising, or working [16].

One further means to understand the functional role of norms is to examine social behaviors that are not governed by norms. For example, consider the problem of moving along a common

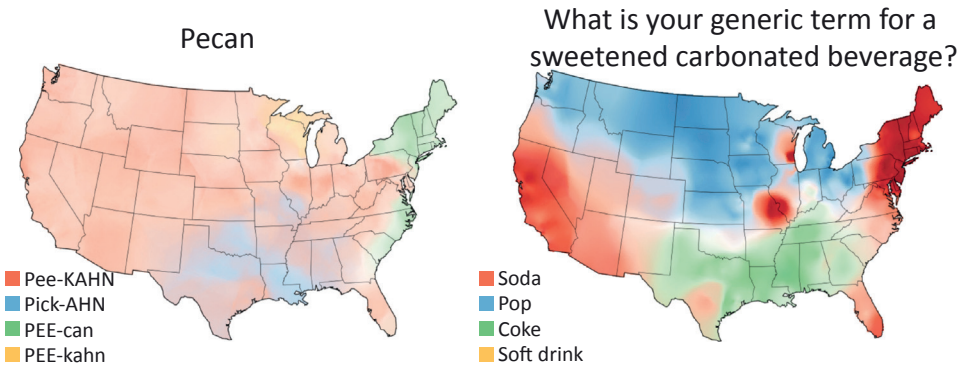
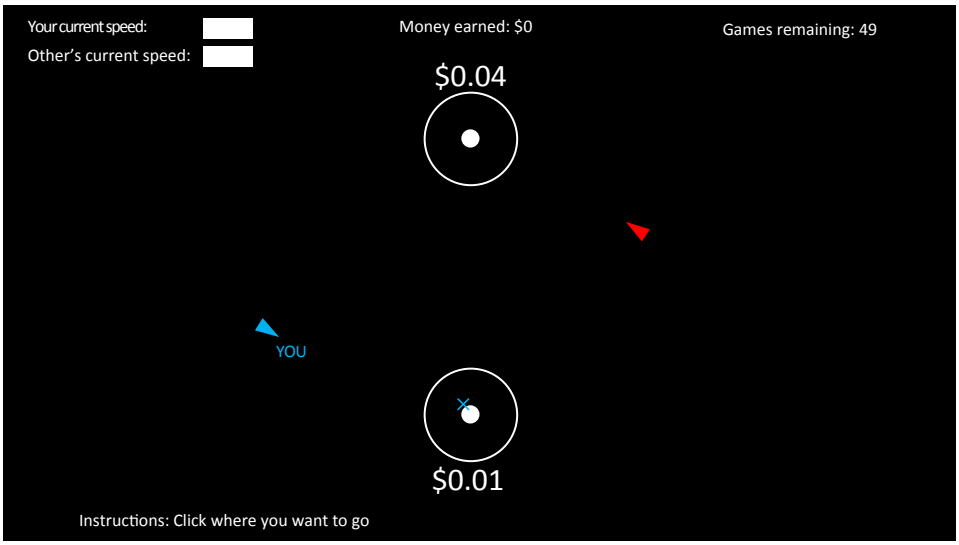


Figure 2. Geographic Distribution of Word Pronunciation and Usage Norms in the USA. The tendency of people to mimic speakers geographically close to them leads to striking regional variation in the pronunciation of words like 'pecan' (top panel) and the word used to refer to sweetened carbonated beverages (bottom panel). Reprinted with permission from [48].

thoroughfare in opposing directions. When we are driving cars, we solve this problem by adhering rigidly to conventions about which side of the road to use. But pedestrians often come to another solution, they just work it out on the fly [17]. This suggests the hypothesis that strong norms only emerge when it is either too inefficient or too costly to dynamically coordinate from scratch during each interaction. Intuitively, miscoordination among cars leads to costly, potentially fatal, vehicular crashes, while pedestrian crashes are awkward at worst. This hypothesis was tested in the lab using a dyadic, real-time coordination game called 'The Battle of the Exes' [18]. Each player was given control of an avatar, which they navigated toward one of two targets with different payoffs. One payoff was larger than the other, but if both players moved to the same target, neither received a bonus (Figure 3). The experiment employed a 2 × 2 design manipulating timing and stakes: in the 'dynamic' condition, players could change their direction at any point during trials and could see their partner's moment-to-moment position, but in the 'ballistic' condition the players select their destinations simultaneously at the beginning of the trial, without subsequent adjustment. In the 'low stakes' condition, there was a small discrepancy between the payoffs (1 cent versus 2 cents), compared with a larger discrepancy (1 cent versus 4 cents) in the 'high stakes' condition.

Across many rounds of this game, some pairs developed that were fair and stable, and led to low rates of miscoordination. For example, players could alternate who gets the high payoff. But such norms were more common under some conditions than others. When the stakes were low, players in the dynamic condition simply relied on moment-to-moment adjustment, just like pedestrians on a crowded street. However, when the stakes were higher, then the dynamic condition developed even more stable norms than the ballistic condition, which was intermediately likely to develop norms regardless of stakes. A similar pattern of results has been obtained from computational agents in a reinforcement learning framework with a sensory-motor control loop [19]. Other factors may influence whether norms will come to govern a behavior beyond the cost of coordination failure and bandwidth of the channel through which



Trends in Cognitive Sciences

Figure 3. Continuous-Time Coordination. Screenshot of Battle of the Exes experiment [18]. Two participants (triangles) move toward one of two payoffs (1 or 4 cents), but if both move to the higher payoff option, then neither gets anything for that round. How would you play this game?

the interaction takes place. For instance, stable norms may be valuable solutions for more frequent behaviors or those that are too complex or cognitively demanding to efficiently handle on the fly, while dynamic coordination may be preferred if agents expect the community to be sparse and short-lived or to lack the organizational infrastructure (e.g., traffic lights for regulating car movements) to sustain norms.

The anticoordinated actions demanded by the Battle of the Exes task also demonstrate that despite the prevalence of conformity and imitation, what matters for navigating social situations is coordinating beliefs and expectations, not necessarily surface behavior. Many groups function better because their members specialize and differentiate their roles [20,21], and move out of sync with one another so as to avoid clashes or redundancy [22]. For example, one study investigated the formation of norms governing roles in a collective task where each player chose a number from 0 to 50 and received feedback about whether the group sum was higher or lower than a target number [23]. Over the course of successive rounds, as the group narrowed in on the target, individuals tended to differentiate themselves into those who were and were not reactive to the feedback. Groups with greater role differentiation were more successful at the task. Norms about roles may also be useful when large groups must harvest the same type or location of limited resources, or in collective search problems where everybody benefits from broadly distributed exploration across individuals [24,25].

Witnessing the Birth of New Norms

Cognitive scientists have employed three primary methods for investigating how norms emerge in a community: naturally occurring datasets describing real-world norms, laboratory investigations of norm creation in small groups of people given simple communication or decision tasks, and computer simulations of interacting agents. There is growing interest in making comparisons across these methods, for example, by validating predictions from computational models with historic records [26,27] or new laboratory data [28–30], or by recreating a naturally observed pattern of norm distribution in simplified laboratory conditions that try to boil down the complex real phenomena to its essence [31]. Here, we describe some of the major themes that have grown out of the cross-pollination of these methods, treating the influences of large-scale group-level processes and individual cognitive processes in turn.

The Influence of Group Processes and Structure on Norm Creation

Long traditions from philosophy, economics, and artificial intelligence have successfully used evolutionary, network-based simulations to account for population-level dynamics of norms and conventions [32–46]. These accounts foreground the importance of group processes and the structure of interactions among individuals (Box 2). For example, in most networks, agents are not uniformly likely to interact with all other agents; they are clustered in local communities or in online ‘echo chambers’. Spatial effects have been investigated thoroughly in simulations [7], and were empirically tested in a recent large-scale experiment [30]. Participants were embedded in a network and randomly matched with their neighbors to play a simple naming game where they received bonus payment only when they and their partner typed the same name for a face. When participants were homogenously matched with all other participants, the whole population converged on a common label, but on other common graph topologies with local structure they tended to get stuck, with local regions of the network using separate labels. This spatial clumpiness of interaction is consistent with the characteristic distribution of word pronunciations and usage that may arise when individuals are more likely to speak to people in their local geographic region [47,48] (Figure 2).

Box 2. Minimal and Sophisticated Agents

Are group-level norms an emergent property of egocentric agents following simple heuristics, or are they supported by more sophisticated cognitive processes and social representations? A primary aim of self-organized, agent-based simulation in the wake of Luc Steels' pioneering language games [118] has been to demonstrate how globally shared signaling systems can arise out of many local interactions of surprisingly minimal agents. A recent synthesis across several such models found three basic ingredients to be necessary: a form of feedback about intended referents as a learning signal, a bias against ambiguity, and a means for forgetting [119]. This research program has elegantly captured a wide variety of collective communication patterns using basic heuristics, from color term categories [120,121] to grammars [46]. Yet minimal agent-based models have also been used to directly argue against the role of more sophisticated social reasoning. After all, if groups of completely egocentric agents can successfully converge, why invoke the additional complexity [122]?

These arguments from minimal agents present a paradox when considered alongside recent computational models of communication and social cognition [74,75,90,123,124], which take local inferences as their primary phenomenon of interest. In order to explain the flexibility and complexity of human social inferences, these models rely on more elaborate cognitive representations and social reasoning processes than their minimal counterparts. By necessity, models simplify factors outside their scope, but why are such different explanations required at different scales? In the synthetic spirit of this review, we propose that there are two ways out of this paradox: ratcheting up the functional demands of individual agents and broadening the scope of target phenomena to be explained. Just as breakthroughs in vision models were driven by considering the challenge of ImageNet's database of millions of photographs organized into thousands of categories rather than MNIST's database of ten hand-drawn digits, progress in minimal multiagent models will be made by considering coordination tasks that increasingly approximate the true computational challenges faced by humans in social contexts [e.g., 125]. At the same time, more cognitively assumptive models initially built to handle the complexities of local interactions should be embedded in larger networks to assess their global properties.

Another critical group-level process shaping norm creation is the formation of persistent institutions that are independent of any particular agents' mind. A community's capacity for creating, enforcing, and revising norms is perhaps its greatest social capital [49]. In fact, many of the most important advances in society can be understood as formally establishing norms (legislature), monitoring for possible norm violations (police), determining whether violations have occurred (courts), and penalizing individuals judged to have violated norms (prisons). While institutions at this scale have been difficult to recreate through laboratory experiments or simulations, naturally occurring datasets abound. Online communities such as Wikipedia have left digital paper trails of explicit discussions concerning group norms [50] and centuries of surviving court records document shifts in norms toward violence [51]. These institutions may reify deep and systematic cultural factors, or meta-norms, such as how restrictive versus loose a society is [52], and catalyze further norm formation by facilitating social interaction processes like preemption, argumentation, negotiation, proposing, straw polling, and voting. These processes, in turn, help groups coordinate on further norms. In contrast to classic game theoretic accounts in which nonbinding promises ('cheap talk') should not have an effect on cooperation, groups in which members are allowed to freely communicate to make proposals, assurances, and promises regarding resource management are more likely to come up with efficient and fair cooperative schemes [53].

While institutions lend stability to norms from generation to generation, intergenerational turnover in the population is a primary mechanism of how norms change [54]. Clearly, a key property of norms is their self-reinforcing stability within a population, but because they are fundamentally grounded in the beliefs of agents, even entrenched norms may shift in part due to interactions involving younger generations who do not yet have these beliefs firmly inculcated [55]. We are often surprised that norms that we think of as long-standing traditions, like a man giving his fiancée a diamond ring when becoming engaged, are actually relatively recent and created as part of a modern-era advertising campaign [56]. A network-wide 'tipping point' dynamic driven by young people locally deciding to adopt a behavior when the proportion of

their neighbors displaying that behavior exceeds some threshold [29] may explain rapid changes in norms for oral sex, smoking indoors, and when texting is appropriate [57]. Mass interventions that shift norm perceptions for a large, or influential, subset of a community can exploit this dynamic for positive societal effect [58]. Even when a behavior is counter to a static norm, evidence that the gradient of that behavior is increasing may motivate change [59].

The effects of intergenerational turnover on norms have been explored experimentally and computationally in 'replacement microsociety' paradigms, where older members of an interacting population are gradually replaced by new learners [60–62]. These experiments thus combine the direct functional pressures of social interaction (discussed in more detail below) with the transmission bottleneck explored by iterated learning paradigms. A familiar example of iterated learning is the children's game of 'telephone', in which a message is passed sequentially along a chain of speakers and listeners. Rather than changing the message randomly, this process has been shown to lead to changes that reflect the inductive biases of learners [63]. Noisy and partial evidence are regularized to fit prior beliefs, so inconsistent norms and complex conventions may gradually grow more systematic and simpler over time. When agents also socially interact within generations, richly structured norms can form. For example, languages tend to become more compositional: instead of holistic systems containing unique signals for each meaning or degenerate systems using a single signal for all meanings, meaningful primitives emerge that can be combined into complex expressions ([64–68], but see [69] for evidence that such systems may also form in the absence of new learners, given other compressibility pressures). This trade-off between complexity and informativity also explains empirical phenomena, such as conventions for color terms across different languages [70,71].

Individual Cognitive Processes That Shape Norms through Local Interaction

The population-level norms that a group ends up establishing are not only shaped by the specifics of the network structure and evolutionary processes external to the members, but also by internal cognitive processes within each member. A fertile area in cognitive science is the attempt to ground population-level phenomena, not in appeals to global equilibria or simple behavioral heuristics, but in the real computational problems faced by agents trying to learn and act in the world (Box 2). Just as expectations about the physical properties of inanimate objects helps an organism navigate the natural world [72,73], well-calibrated mental models that accurately predict behaviors of other agents help us navigate the social world [74,75]. If we assume there is some latent norm in place governing others' behavior, but initially have uncertainty over what it is, we can attempt to infer it from observing other agents. Thus, complex norms and conventions may get off the ground through social reasoning. In our search for social structure, we create it [10,76–78].

Sometimes the cognitive processes leading to norms are straightforward and driven only by a functional need to synchronize our raw behavior [79]. For example, when partners align to one another's word choices [80], syntax [81], body postures [82], or even informational complexity [83], they may be directly adapting to the surface statistics of their partner [84]. In real-time conversation, people depend on feedback through backchannels to dynamically repair their utterances as they attempt to coordinate on an intended meaning [85]. When observing descriptive norms, agents may simply use prevalence as a heuristic cue to which behaviors are advantageous [24,86].

In other cases, the connection between individual cognition and norms is subtler and may involve more sophisticated social representations. While traditional definitions in philosophy

have often stipulated an idealized property called common knowledge (requiring infinitely recursive beliefs about others' beliefs) in practice, only a couple of levels may be required. The appropriate level of reasoning may itself be a norm [87]. If agents are motivated to do what others around them are doing to build relationships, obtain social approval, manage others' impressions, or signal social identity [88], they must maintain some latent representation of others' expectations. In communication, listeners generally expect that speakers are attempting to balance informativity and parsimony: intending to tell the truth and give information but no more than is required by the situation [89–91]. In other words, we not only actively attempt to have our meanings understood, we charitably assume that others are trying to do so as well. This allows for remarkable contextual flexibility in making *ad hoc* pragmatic inferences that go beyond literal meanings, for instance in understanding hyperbole or irony [92]. Even when a partner uses a word in a completely novel, unconventional sense, it is often easy to infer what is intended and accommodate it [93]. Indeed, interacting dyads will dynamically and jointly switch on a trial-by-trial basis their interpretation of a simple signal based on the environment and task constraints [94].

Repeated reference games provide a generative paradigm for studying the cognitive processes involved in rapid, *ad hoc* convention formation among dyads. In these experiments, a designated 'sender' and 'receiver' communicate about a set of objects; the sender attempts to convey the identity of one of these objects to a receiver, and the receiver attempts to use their messages to select the intended one [95,96]. One recent version of this is a Pictionary-like game in which the sender is tasked with getting the receiver to guess a word [97–99] or piece of music [100] based on the sender's graphical sketching. Example sketches resulting from different conditions of the word-guessing game are shown in Figure 4.

We highlight three key computational challenges this task poses for agents, which models of norm creation ought to capture. First, at the outset of the game, both players must harness prior population-level expectations from their knowledge of the medium to successfully communicate difficult-to-express words to a completely novel partner. In linguistic games, this typically involves using longer, more specific descriptions than we would provide just for ourselves [101] and inferring shared cultural background that can be exploited for efficiency [102]; in sketching games, it involves highly detailed drawings making use of iconicity, or expectations about how a particular set of strokes will be perceived to resemble things in the world. Second, through rapid learning on earlier rounds, agents must acquire expectations allowing them to move from universally understood iconic messages or sketches to more efficient but idiosyncratic 'symbolic' representations [97,103–105]. As a result, participants not directly involved in the dyadic interaction are not good at correctly interpreting later messages [106]. These local conventions depend on the local context [107] and the feedback channel [97,108,109]. When senders do not receive feedback from receivers on whether their messages have been understood, as shown in Figure 4A, then they do not become simpler over repetitions. Third, the local expectations about meanings that the agent learned through interaction must be represented as partner-specific (if the receiver is swapped out with a novel partner, senders revert back to their initial messages [96,110]) but, with sufficient consistency of partners, may generalize to global expectations. For example, when each player was embedded in a small, fully connected network and played a repeated reference game with all others in turn, agents were more and more willing to start a new interaction with the convention they had converged on in the previous interaction, thus leading the whole population to converge on a shared prior [99,111].

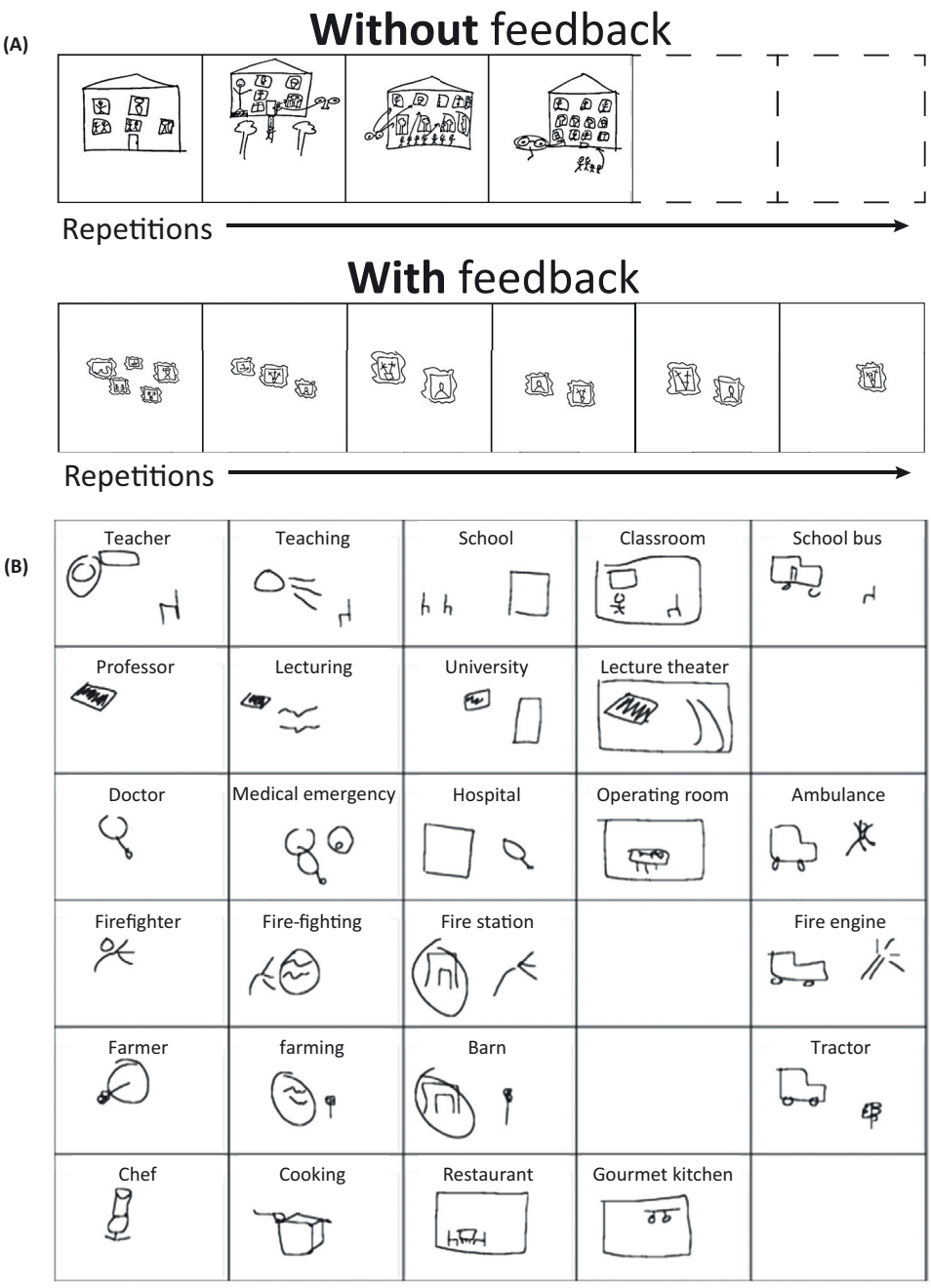


Figure 4. Sample Sketches in Laboratory Tasks on Graphical Communication. Interacting pairs rapidly coordinate on efficient and systematic conventions for referring to concepts using drawings. (A) Social interaction is critical for convention formation [97]. Example drawings indicating 'art gallery' quickly become more efficient when the receiver is allowed to provide graphical feedback, but persist in complexity without real-time feedback (i.e., when the receiver sees and identifies the drawings offline). (B) Repeated interaction results in the formation of graphical conventions displaying systematicity [104]. Taking one pair as a case study, each of their drawings related to university education feature a filled in diamond, and most of their drawings for 'activities' (in the second column) have parallel, squiggly lines. Some cells are empty because the corresponding concept was not used in the experiment. Reprinted with permission from the authors.

Concluding Remarks

The norms that emerge in a community will be shaped by the cognitive processes within each individual, operating in local dyadic interactions, as well as the broader population-level infrastructure of existing norms in which these interactions are embedded. Together, these influences characterize a dynamic, multidirectional process for norm evolution. Understanding how processes operating across different temporal and spatial scales interact will be pivotal for being able to predict and control the norms that shape society (see Outstanding Questions). Individual agents are trying to learn about others' underlying beliefs and behaviors to more successfully navigate their social world, hence their initial learning is regularized by their priors and strengthened by assumptions about others' intentions. When multiple agents in a population all expect there to be a regularity and attempt to learn it from one another, a norm or convention emerges. It will then go on to facilitate and constrain further interactions among the individuals at a population level. One effect of norms will be to modify social networks and streamline communication channels. As we have already seen, these affected social networks and communication channels will, in turn, guide future norm creation. Norms are both the consequence and facilitator of social interaction.

Outstanding Questions

How does the modality (pictures, words, sounds) through which signals are exchanged influence the initial messages as well as the resulting conventions that are formed? What kinds of signals are most useful for communicating about different topics and which are most likely to foster frequently beneficial properties of a communication system, such as syntax, systematicity, efficiency, and compositionality?

How well can laboratory-based methods for investigating the emergence of norms emulate naturally occurring processes of norm creation? What kinds of norm-creation processes are best studied in the laboratory versus by analyzing real-world interactions? Are laboratory experiments blinding researchers to socially important processes that require years or decades, rich communication, or pre-existing infrastructural scaffolding in order to unfold?

How do unidirectional dynamics of cultural transmission acting over long time scales (as explored in iterated experiments) interact with shorter time scales of bidirectional coordination (as explored in dyadic experiments)?

Can we develop a formal model integrating previously established norms, individual psychological constraints, and technological supports for social interaction to predict norms that will arise in different communities, and aid in the formation of improved norms?

How do some norms that begin as purely descriptive or conventional begin to take on additional prescriptive force? How is prescriptive force intertwined with social group membership?

How are neural systems integrated to support rapid partner-specific convention formation and selective generalization to novel partners and contexts?

Acknowledgements

We thank Simon DeDeo, Edgar Jose Andrade Lotero, and Paul Smaldino for their helpful comments. The first author was supported by the Stanford University Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

1. Appiah, K.A. (2011) *The Honor Code: How Moral Revolutions Happen,* WW Norton & Company
2. Papachristos, A.V. (2009) Murder by structure: dominance relations and the social structure of gang homicide. *AJS* 115, 74–128
3. Zhang, J. (2010) The sound of silence: observational learning in the U.S. kidney market. *Market. Sci.* 29, 315–335
4. Borsari, B. and Carey, K.B. (2003) Descriptive and injunctive norms in college drinking: a meta-analytic integration. *J. Stud. Alcohol* 64, 331–341
5. O'Callaghan, F.V. and Nausbaum, S. (2006) Predicting bicycle helmet wearing intentions and behavior among adolescents. *J. Safety Res.* 37, 425–431
6. Young, H.P. (2015) The evolution of social norms. *Annu. Rev. Econ.* 7, 359–387
7. Baronchelli, A. (2018) The emergence of consensus: a primer. *R. Soc. Open Sci.* 5, 172189
8. Ehrlich, P.R. and Levin, S.A. (2005) The evolution of norms. *PLoS Biol.* 3, e194
9. Lewis, D. (1969) *Convention,* Harvard University Press
10. Bicchieri, C. (2006) *The Grammar of Society,* Cambridge University Press
11. Ullmann-Margalit, E. (1977) *The Emergence of Norms,* Clarendon Press
12. Hume, D. (1740) . In *A Treatise of Human Nature* (. In *A Treatise of Human Nature* 3rd edn (Selby-Brigge, L.A. and Nidditch, P., eds), Clarendon Press
13. Goldberg, A.E. (2006) *Constructions at Work: The Nature of Generalization in Language,* Oxford University Press
14. Hornstein, N. *et al.* eds. (2018) *Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics* (Vol. 129), Walter de Gruyter GmbH & Co KG
15. Frey, S. and Goldstone, R.L. (2013) Cyclic game dynamics driven by iterated reasoning. *PLoS One* 8, e56416
16. Ashworth, T. (1980) *Trench Warfare 1914–1918: The Live and Let Live System,* Macmillan
17. Helbing, D. and Molnar, P. (1995) Social force model for pedestrian dynamics. *Phys. Rev.* E51, 4282–4286
18. Hawkins, R.X.D. and Goldstone, R.L. (2016) The formation of social conventions in real-time environments. *PLoS One* 11, e0151670
19. Freire, I.T. *et al.* (2018) Modeling the formation of social conventions in multi-agent populations. *arXiv* 1802.06108
20. Sloman, S. and Fernbach, P. (2017) *The Knowledge Illusion: Why We Never Think Alone,* Riverhead Books
21. Smaldino, P.E. (2014) The cultural evolution of emergent group-level traits. *Behav. Brain Sci.* 37, 243–295
22. Nalepka, P. *et al.* (2017) Herd those sheep: emergent multiagent coordination and behavioral-mode switching. *Psychol. Sci.* 28, 630–650
23. Roberts, M.E. and Goldstone, R.L. (2011) Adaptive group coordination and role differentiation. *PLoS One* 6, 1–8
24. Hills, T.T. *et al.* (2015) Exploration versus exploitation in space, mind, and society. *Trends Cogn. Sci.* 19, 46–54
25. Goldstone, R.L. *et al.* (2013) Learning along with others. *Psychol. Learn. Motiv.* 58, 1–45
26. Bowles, S. and Choi, J.K. (2013) Coevolution of farming and private property during the early Holocene. *Proc. Natl. Acad. Sci. U. S. A.* 110, 8830–8835
27. Conte, C. *et al.* (2013) *Minding Norms: Mechanisms and Dynamics of Social Order in Agent Societies,* Oxford University Press
28. Barkoczi, D. and Galesic, M. (2016) Social learning strategies modify the effect of network structure on group performance. *Nat. Commun.* 7, 13109
29. Centola, D. *et al.* (2018) Experimental evidence for tipping points in social convention. *Science* 360, 1116–1119
30. Centola, D. and Baronchelli, A. (2015) The spontaneous emergence of conventions: an experimental study of cultural evolution. *Proc. Natl. Acad. Sci. U. S. A.* 112, 1989–1994
31. Kempe, M. and Mesoudi, A. (2014) Experimental and theoretical models of human cultural evolution. *Wiley Interdiscip. Rev. Cogn. Sci.* 5, 317–326
32. Skyrms, B. (2010) *Signals: Evolution, Learning, and Information,* Oxford University Press

33. Sugden, R. (1989) Spontaneous order. *J. Econ. Perspect.* 3, 85–97
34. Young, H.P. (1993) The evolution of conventions. *Econometrica* 61, 57–84
35. Axelrod, R. (1986) An evolutionary approach to norms. *Am. Pol. Sci. Rev.* 80, 1095–1111
36. Ellison, G. (1993) Learning, local interaction, and coordination. *Econometrica* 61, 1047–1071
37. Efferson, C. *et al.* (2008) Conformists and mavericks: the empirics of frequency-dependent cultural transmission. *Evol. Hum. Behav.* 29, 56–64
38. Elsenbroich, C. and Gilbert, N. (2013) *Modelling Norms,* Springer Science & Business Media
39. Binmore, K. and Samuelson, L. (2006) The evolution of focal points. *Games Econ. Behav.* 55, 21–42
40. Macy, M.W. and Skvoretz, J. (1998) The evolution of trust and cooperation between strangers: a computational model. *Am. Sociol. Rev.* 63, 638–660
41. Van Rooy, R. (2004) Evolution of conventional meaning and conversational principles. *Synthese* 139, 331–366
42. Shoham, Y. and Tennenholtz, M. (1997) On the emergence of social conventions: modeling, analysis, and simulations. *Artif. Intell.* 94, 139–166
43. Delgado, J. (2002) Emergence of social conventions in complex networks. *Artif. Intell.* 141, 171–185
44. Mühlenbernd, R. (2011) Learning with neighbours. *Synthese* 183, 87–109
45. Bruner, J. *et al.* (2018) David Lewis in the lab: experimental results on the emergence of meaning. *Synthese* 195, 603–621
46. Steels, L., ed. (2012). *Experiments in Cultural Language Evolution* (Vol. 3), John Benjamins Publishing
47. Sneller, B. and Roberts, G. (2018) Why some behaviors spread while others don't: a laboratory simulation of dialect contact. *Cognition* 170, 298–311
48. Katz, J. (2016) *Speaking American: How Y'all, Youse, and You Guys Talk: A Visual Guide,* Houghton Mifflin Harcourt
49. Ostrom, E. (2000) Collective action and the evolution of social norms. *J. Econ. Perspect.* 14, 137–158
50. Heaberlin, B. and DeDeo, S. (2016) The evolution of Wikipedia's norm network. *Future Internet* 8, 14
51. Klingenstein, S. *et al.* (2014) The civilizing process in London's Old Bailey. *Proc. Natl. Acad. Sci. U. S. A.* 111, 201405984
52. Gelfand, M. *et al.* (2011) Differences between tight and loose cultures: a 33-nation study. *Science* 332, 1100–1104
53. Ostrom, E. (2006) The value-added of laboratory experiments for the study of institutions and common-pool resources. *J. Econ. Behav. Organ.* 61, 149–163
54. Labov, W. (2007) Transmission and diffusion. *Language* 83, 344–387
55. Danescu-Niculescu-Mizil, C. *et al.* (2013) No country for old members: user lifecycle and linguist change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 307–318, Rio de Janeiro, Brazil
56. Newbury, C.W. (1990) *The Diamond Ring: Business, Politics, and Precious Stones in South Africa, 1867-1947,* Clarendon Press
57. Nyborg, K. *et al.* (2016) Social norms as solutions. *Science* 354, 42–43
58. Tankard, M.E. and Paluck, E.L. (2016) Norm perception as a vehicle for social change. *Soc. Issues Policy Rev.* 10, 181–211
59. Sparkman, G. and Walton, G.M. (2017) Dynamic norms promote sustainable behavior, even if it is counternormative. *Psychol. Sci.* 28, 1663–1674
60. Mesoudi, A. and Whiten, A. (2008) The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 3489–3501
61. Acerbi, A. and Parisi, D. (2006) Cultural transmission between and within generations. *J. Artif. Soc. Soc. Simul.* 9, 1–9
62. Caldwell, C.A. and Smith, K. (2012) Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS One* 7, e43807
63. Kirby, S. *et al.* (2014) Iterated learning and the evolution of language. *Curr. Opin. Neurobiol.* 28, 108–114
64. Kirby, S. *et al.* (2015) Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141, 87–102
65. Vogt, P. (2005) The emergence of compositional structures in perceptually grounded language games. *Artif. Intell.* 167, 206–242
66. Carr, J.W. *et al.* (2017) The cultural evolution of structured languages in an open-ended, continuous world. *Cogn. Sci.* 41, 892–923
67. Kirby, S. *et al.* (2007) Innateness and culture in the evolution of language. *Proc. Natl. Acad. Sci. U. S. A.* 104, 5241–5245
68. Tamariz, M. and Kirby, S. (2016) The cultural evolution of language. *Curr. Opin. Psychol.* 8, 37–43
69. Raviv, L. *et al.* (2018) Compositional structure can emerge without generational transmission. *Cognition* 182, 151–164
70. Gibson, E. *et al.* (2017) Color naming across languages reflects color use. *Proc. Natl. Acad. Sci. U. S. A.* 114, 10785–10790
71. Zaslavsky, N. *et al.* (2018) Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci. U. S. A.* 115, 7937–7942
72. Ullman, T.D. *et al.* (2017) Mind games: game engines as an architecture for intuitive physics. *Trends Cogn. Sci.* 21, 649–665
73. Spelke, E.S. *et al.* (1992) Origins of knowledge. *Psychol. Rev.* 99, 605
74. Baker, C.L. *et al.* (2017) Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* 1, 0064
75. Jara-Ettinger, J. *et al.* (2016) The naïve utility calculus: computational principles underlying commonsense psychology. *Trends Cogn. Sci.* 20, 589–604
76. Hawkins, R.X.D. *et al.* (2017) Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Cognitive Science Society
77. Muldoon, R. *et al.* (2014) Why are there descriptive norms? Because we looked for them. *Synthese* 191, 4409–4429
78. Ho, M.K. *et al.* (2016) *Feature-Based Joint Planning and Norm Learning in Collaborative Games,* Cognitive Science Society
79. Fusaroli, R. *et al.* (2012) Coming to terms: quantifying the benefits of linguistic coordination. *Psychol. Sci.* 23, 931–939
80. Louwerse, M.M. *et al.* (2012) Behavior matching in multimodal communication is synchronized. *Cogn. Sci.* 36, 1404–1426
81. Levelt, W.J.M. and Kelter, S. (1982) Surface form and memory in question answering. *Cognit. Psychol.* 14, 78–106
82. Lakin, J.L. and Chartrand, T.L. (2003) Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychol. Sci.* 14, 334–339
83. Abney, D.H. *et al.* (2014) Complexity matching in dyadic conversation. *J. Exp. Psychol. Gen.* 143, 2304–2315
84. Garrod, S. and Pickering, M.J. (2004) Why is conversation so easy? *Trends Cogn. Sci.* 8, 8–11
85. Dingemanse, M. *et al.* (2015) Universal principles in the repair of communication problems. *PLoS One* 10, e0136100
86. Rendell, L. *et al.* (2011) Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends Cogn. Sci.* 15, 68–76
87. Frey, S. and Goldstone, R.L. (2018) Cognitive mechanisms for human flocking dynamics. *J. Comput. Soc. Sci.* 1, 349–375
88. Cialdini, R.B., and Trost, M.R. (1998). Social influence: social norms, conformity and compliance. In *The Handbook of Social Psychology* (Vols. 1–2, 4th edn) (Gilbert, D.T. *et al.*, eds), pp. 151–192, Wiley
89. Grice, H.P. (1975) Logic and conversation. In *Syntax and Semantics* (Vol. 3, Speech Acts) (Cole, P. and Morgan, J.L., eds), pp. 41–58, Academic Press

90. Goodman, N.D. and Frank, M.C. (2016) Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* 20, 818–829

91. Franke, M. and Jäger, G. (2016) Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Z. Sprachwissenschaft* 35, 3–44

92. Kao, J.T. *et al.* (2014) Nonliteral understanding of number words. *Proc. Natl. Acad. Sci. U. S. A.* 111, 12002–12007

93. Clark, E.V. and Clark, H.H. (1979) When nouns surface as verbs. *Language* 55, 767–811

94. Misyak, J. *et al.* (2016) Instantaneous conventions: the emergence of flexible communicative signals. *Psychol. Sci.* 27, 1550–1561

95. Krauss, R.M. and Weinheimer, S. (1964) Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychon. Sci.* 1, 113–114

96. Wilkes-Gibbs, D. and Clark, H.H. (1992) Coordinating beliefs in conversation. *J. Mem. Lang.* 31, 183–194

97. Garrod, S. *et al.* (2007) Foundations of representation: where might graphical symbol systems come from? *Cogn. Sci.* 31, 961–987

98. Garrod, S. *et al.* (2010) Can iterated learning explain the emergence of graphical symbols? *Interect. Stud.* 11, 33–50

99. Fay, N. *et al.* (2010) The interactive evolution of human communication systems. *Cogn. Sci.* 34, 1–36

100. Healy, P.G.T. *et al.* (2007) Graphical language games: interactional constraints on representational form. *Cogn. Sci.* 31, 285–309

101. Fussell, S.R. and Krauss, R.M. (1989) The effects of intended audience on message production and comprehension: reference in a common ground framework. *J. Exp. Soc. Psychol.* 25, 203–219

102. Isaacs, E.A. and Clark, H.H. (1987) References in conversation between experts and novices. *J. Exp. Psychol. Gen.* 116, 26

103. Dingemanse, M. *et al.* (2015) Arbitrariness, iconicity, and systematicity in language. *Trends Cogn. Sci.* 19, 603–615

104. Theisen, C.A. *et al.* (2010) Systematicity and arbitrariness in novel communication systems. *Interact. Stud.* 11, 14–32

105. Little, H. *et al.* (2017) Signal dimensionality and the emergence of combinatorial structure. *Cognition* 168, 1–15

106. Schober, M.F. and Clark, H.H. (1989) Understanding by addressees and overhearers. *Cognit. Psychol.* 21, 211–232

107. Hawkins, R.X.D. *et al.* (2018) Emerging abstractions: lexical conventions are shaped by communicative context. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Cognitive Science Society

108. Hupet, M. and Chantraine, Y. (1992) Changes in repeated references: collaboration or repetition effects? *J. Psycholinguist. Res.* 21, 485–496

109. Krauss, R.M. and Weinheimer, S. (1966) Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *J. Pers. Soc. Psychol.* 4, 343

110. Brennan, S.E. and Clark, H.H. (1996) Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482

111. Garrod, S. and Doherty, G. (1994) Conversation, coordination and convention—an empirical investigation of how groups establish linguistic conventions. *Cognition* 53, 181–215

112. Turiel, E. (1983) *The Development of Social Knowledge: Morality and Convention,* Cambridge University Press

113. Cialdini, R. *et al.* (1990) A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. *Adv. Exp. Soc. Psychol.* 24, 201–234

114. Brennan, G. *et al.* (2013) *Explaining Norms,* Oxford University Press

115. Southwood, N. and Eriksson, L. (2011) Norms and conventions. *Philos. Explor.* 14, 195–217

116. Fehr, E. and Fischbacher, U. (2004) Social norms and human cooperation. *Trends Cogn. Sci.* 8, 185–190

117. Rakoczy, H. *et al.* (2008) The sources of normativity: young children's awareness of the normative structure of games. *Dev. Psychol.* 44, 875

118. Steels, L. (1995) A self-organizing spatial vocabulary. *Artif. Life* 2, 319–332

119. Spike, M. *et al.* (2017) Minimal requirements for the emergence of learned signaling. *Cogn. Sci.* 41, 623–658

120. Baronchelli, A. *et al.* (2010) Modeling the emergence of universality in color naming patterns. *Proc. Natl. Acad. Sci. U. S. A.* 107, 2403–2407

121. Loreto, V. *et al.* (2012) On the origin of the hierarchy of color names. *Proc. Natl. Acad. Sci. U. S. A.* 109, 6819–6824

122. Barr, D.J. (2004) Establishing conventional communication systems: is common knowledge necessary? *Cogn. Sci.* 28, 937–962

123. Rabinowitz, N.C. *et al.* (2018) Machine theory of mind. *arXiv* 1802.07740

124. Shafto, P. *et al.* (2014) A rational account of pedagogical reasoning: teaching by, and learning from examples. *Cogn. Psychol.* 71, 55–89

125. Lazaridou, A. *et al.* (2018) Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv* 1804.03984