# Statistical learning of social signals and its implications for the social brain hypothesis

Hjalmar K. Turesson & Asif A. Ghazanfar
Princeton University, USA

The social brain hypothesis implies that humans and other primates evolved "modules" for representing social knowledge. Alternatively, no such cognitive specializations are needed because social knowledge is already present in the world — we can simply monitor the dynamics of social interactions. Given the latter idea, what mechanism could account for coalition formation? We propose that statistical learning can provide a mechanism for fast and implicit learning of social signals. Using human participants, we compared learning of social signals with arbitrary signals. We found that learning of social signals was no better than learning of arbitrary signals. While coupling faces and voices led to parallel learning, the same was true for arbitrary shapes and sounds. However, coupling versus uncoupling social signals with arbitrary signals revealed that faces and voices are treated with perceptual priority. Overall, our data suggest that statistical learning is a viable domain-general mechanism for learning social group structure.

**Keywords:** social brain; embodied cognition; distributed cognition; situated cognition; multisensory; audiovisual speech; crossmodal; multimodal

## 1. Introduction

The environmental challenges facing primates were no more taxing than those faced by other species, yet the brains of monkeys, apes and humans are bigger than expected. The "social brain hypothesis" attempts to explain this by suggesting that the *social* environment provided unique cognitive challenges (Humphrey 1976; Jolly 1966). As a consequence of living permanently in social groups with local competition for scarce resources, the pressure for primates to evolve an ability to outwit other group members was exceptionally strong. The socially smart animals would enjoy increased survival and reproductive success. Support for this hypothesis is the strong positive correlation between neocortex size and social group size across primate species (Dunbar 1992, 1995, 1998; Kudo & Dunbar 2001), and even among social carnivores (Perez-Barberia, Shultz, & Dunbar 2007).

*Interaction Studies* 12:3 (2011), 397–417. DOI 10.1075/is.12.3.02tur
ISSN 1572–0373 / E-ISSN 1572–0381 © John Benjamins Publishing Company

The social brain hypothesis presents a view of primates as biologically-prepared for social signals such as faces and voices as well as for forms of social engagements that require mental representations of abstract concepts like family relations and alliances in order to negotiate the social landscape (Cheney & Seyfarth 2007; Ghazanfar & Santos 2004; Kurzban, Tooby, & Cosmides 2001). An alternative hypothesis suggests that individuals do not need to hold abstract concepts of family relations and alliances 'in mind' because they can assess circumstances by directly monitoring what is happening around them (Barrett & Henzi 2005; Barrett, Henzi, & Rendall 2007; Barrett & Rendall 2009; Johnson 2001). According to this view, the active perception of on-going spatial and temporal structure of interacting primates within a social group obviates the need for high level processing involving mental representations. Individuals can use this on-going structure as an accurate and always up-to-date model, allowing for more efficient action selection and execution (Clark 1997; Pfeifer & Scheier 1999).

Among the characteristics of an integrated social organization, as distinguished from a random aggregation, is cohesion — a tendency of the members to remain together (Eisenberg 1965). One pattern recognition mechanism that could be used to actively and rapidly track such cohesion in the social environment is statistical learning. Statistical learning refers to the capacity to segment the sensory environment, based on probabilistically-defined patterns, without intention or awareness. This learning mechanism provides us with the ability to infer, without instruction, which features of an initially unstructured sensory input belong together. The ability to learn statistical regularities in sensory input is frequently explored with experiments that test how subjects learn the association patterns of unitary elements defined by the statistics of element co-occurrence (Fiser & Aslin 2001, 2002a; J.R. Saffran, Aslin, & Newport 1996; J.R. Saffran, Johnson, Aslin, & Newport 1999; Turk-Browne, Jung, & Scholl 2005). The overall consensus seems to be that statistical learning is mediated by a general mechanism that operates over several types of sensory patterns, and across the auditory, visual, and tactile modalities (Conway & Christiansen 2005, 2006; Fiser & Aslin 2001; J.R. Saffran, et al. 1996). In the context of primate social cognition, such a learning mechanism could be used to rapidly and flexibly group individuals into cliques based on their spatiotemporal patterns of association. In such a scenario, the *faces* and *voices* of individuals would likely be important sensory elements.

Several lines of evidence suggest that faces are a special class of stimuli, a class that relies on different neural systems (McKone, Kanwisher, & Duchaine 2007; Tsao, Cadieu, & Livingstone 2010) and perceptual strategies (Bruce & Young 1986; Sugita 2008) when compared to object recognition, and exhibits a unique developmental trajectory (Scherf, Behrmann, Humphreys, & Luna 2007). Faces are also typically accompanied by voices which are themselves thought to be processed by specialized neural circuits (Belin, Fecteau, & Bedard 2004) and also develop with a unique trajectory in humans (Werker & Tees 1984). In the following study, we investigated whether the statistical learning by human subjects of the temporal structure of triplets of faces, voices or their combination was possible and, if so, whether it was different from the learning of synthetic visual objects and sounds. We focused on triplets because conversational cliques composed of more than five people are rare, and clique size reaches an asymptote at three individuals (Dunbar, Duncan, & Nettle 1995). There are four potential, mutually exclusive outcomes to our study. First, since faces and voices are among the most salient features of the human environment, it is possible that that statistical learning is better for these signals than they would be for artificial stimuli. A second possibility is that, because faces and voices seem to be processed by specialized perceptual and neural strategies (different from those used for generic object and sound recognition), the statistical learning paradigm may not operate, or operate sub-optimally, for such signals. A third possibility is that faces and voices and artificial objects and sounds are learned equally well under a statistical learning paradigm. The fourth and final possibility is that the typical, everyday crossmodal association of human faces and voices leads to differential learning when compared to associations of objects and sounds. We executed a series of experiments to investigate these putative differences between the statistical learning of faces and voices versus arbitrary shapes and sounds.

## 2. Experiment 1: Sequential grouping of unimodal visual and auditory signals

In Experiment 1, we assessed the statistical learning of *sequences* of faces versus shapes and voices versus synthesizer sounds, using a well-established paradigm (Fiser & Aslin 2001, 2002a; J.R. Saffran, et al. 1996; Turk-Browne, et al. 2005). The experiment had three possible outcomes for each of the two sensory modalities: faces (or voices) are learned better than shapes (or synthesizer sounds), shapes (or synthesizer sounds) are learned better than faces (or voices), or the two stimulus classes within a modality are learned equally well.

### 2.1 Methods

*Subjects.* Two separate groups of subjects were tested, one for the visual modality and the other for the auditory modality. There were twenty naïve subjects per group and subjects participated in two counterbalanced conditions. All subjects were undergraduate students at Princeton University without overt visual or hearing disabilities. Previous studies have demonstrated statistical learning in all ages ranging from infants (Fiser & Aslin 2002b), young adults, to older adults with an age of around 70 years (Howard, Howard, Dennis, & Kelly 2008). Thus age does not appear to be of critical importance. Handedness was not noted, and since the test phase was self-paced and we did not measure reaction times, and thus we do not expect handedness to influence the results in any way.

*Stimuli.* We used two sets of visual elements: faces and shapes. Faces were gray-scale with 6.5–7.0 cm height and 5.5–7.0 cm width (Figure 1). Face stimuli came from NimStim, www.macbrain.org/faces. They consisted of 6 female and 6 male faces without explicit emotional expressions. Twelve shapes were created, modeled after the shapes used by Fiser and Aslin (Fiser & Aslin 2001). The shapes were white on a black background with 2.5–4.0 cm height and 1.5–4.0 cm width. Visual stimuli were presented on a 17-inch Dell Trinitron display at a comfortable viewing distance (40–90 cm) that readily gives access to the keyboard for responses during the test phase. For the auditory stimuli, we used voices and synthesizer sounds. For the voice elements, twelve 500-millisecond long segments of natural speech were edited with the only requirement that they should not be recognizable as words, as judged by the authors. This allowed for them to be composed of one or more syllables, or cross syllable boundaries. Each voice element originated from a different speaker. The speech was taken from the Santa Barbara Corpus of Spoken American English. The synthesizer sounds (hereafter, *syn-sounds*) were twelve 500-ms long segments, lacking any biological salience, but yet, like the voice elements, spectrotemporally complex. All auditory stimuli were played at 75 dB as measured from 60 cm distance from the screen.

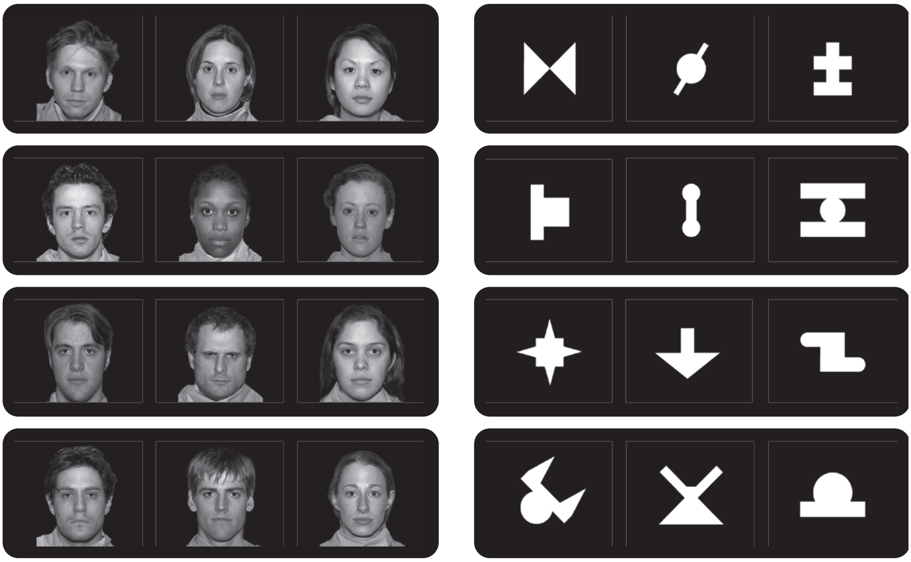

**Figure 1.** Left, example triplets of the 12 faces, and, right, example triplets of the 12 arbitrary shapes used as visual stimuli. Note, that for each subject the elements making up a triplet was randomized

*Design.* We used a sequential design, dividing each stimulus set into four short sequences of three elements, or *base-triplets*. To generate the familiarization streams we arranged the four base-triplets in a 96 base-triplet long randomly ordered sequence, with the constraint that no consecutive occurrences of the same base-triplet were allowed.

*Procedure.* The familiarization phase lasted for a total period of approximately 2 minutes and 50 seconds. Elements were presented for 500-ms, and separated by 100-ms. Subjects were asked to pay attention to the sequence so that they would be able to answer some questions after the familiarization phase. The familiarization phase was followed by a test phase that consisted of a temporal two-alternative forced choice (2AFC) test with 8 trials. During the test phase, a base-triplet and a cross-triplet were shown or heard sequentially, separated by 1 second. *Cross-triplets* were constructed as semi-random three-element sequences that never appeared in that particular order during the familiarization phase; they were rearranged mixtures of base-triplets (i.e. if two base triplets are ABC and DEF, then EDA is a possible cross-triplet). Since, the number of possible cross-triplets is greater than the number of base-triplets, we randomly selected cross-triplets for each subjects. Each pair of base-triplet and cross-triplet was shown twice, in semi-random and counterbalanced order. Subjects reported which of the two test patterns they found most familiar through pressing the "Z" or "M" keys on a keyboard. For every subject, the order of test triplets was randomized. Once a subject completed the experiment with one stimulus class (e.g. shapes), she/he did the experiment again, but with the other stimulus class (e.g. faces). The same was true for the subjects performing the auditory conditions. The order of conditions was counterbalanced.

## 2.2 Results and discussion

Subjects were able to learn triplet sequences of both faces and shapes (Figure 2A), and they were able to learn them equally well (no main effect for stimulus class: ($F(1, 18) = 1.24$, $p = 0.28$). Mean performance for faces was 62.5% (chance performance is 50% for this and all subsequent tests below) ($t(19) = 2.70$, $p < 0.05$) and 67.5% for shapes ($t(19)$-3.44, $p < 0.01$). There was no effect of condition order, ($F(1, 18) = 2.77$, $p = 0.11$), suggesting subjects' participation in two consequent experiments did not influence learning. The same pattern held for the learning of sequences of voice and syn-sound elements (Figure 2B). Subjects learned the two types of elements equally well (no main effect for stimulus class: ($F(1, 18) = 0.23$, $p = 0.79$). Mean performance for voices was 62.5% ($t(19) = 2.70$, $p < 0.05$) and 60.6% for syn-sounds ($t(19) = 2.43$, $p < 0.05$). There was no influence of condition order on learning ($F(1, 18) = 0.013$, $p = 0.91$). Furthermore, there were no performance differences across modalities. The mean performance for visual elements (faces and shapes combined) was 65.0%, while performance for auditory elements (voices and syn-sounds combined) was 61.6% (t(38) = 0.67, p = 0.51). The 60–65% performance we observed is comparable to what has been reported earlier for familiarization streams with a similar presentation rate (Turk-Browne, et al. 2005).

These results suggest that subjects were able to implicitly learn statistical regularities equally well for faces and shapes and for voices and syn-sounds (Figure 2). Thus, social signals are not better or worse in this type of learning, despite their special status in perception. One key distinction between faces and voices versus shapes and syn-sounds is that the former are often reliably associated with one another. That is, specific faces are associated with specific voices, whereas there is no such relationship between the arbitrary shapes and syn-sounds. In next three experiments, we test the influence of this association (or lack thereof) in the statistical learning of triplet sequences.

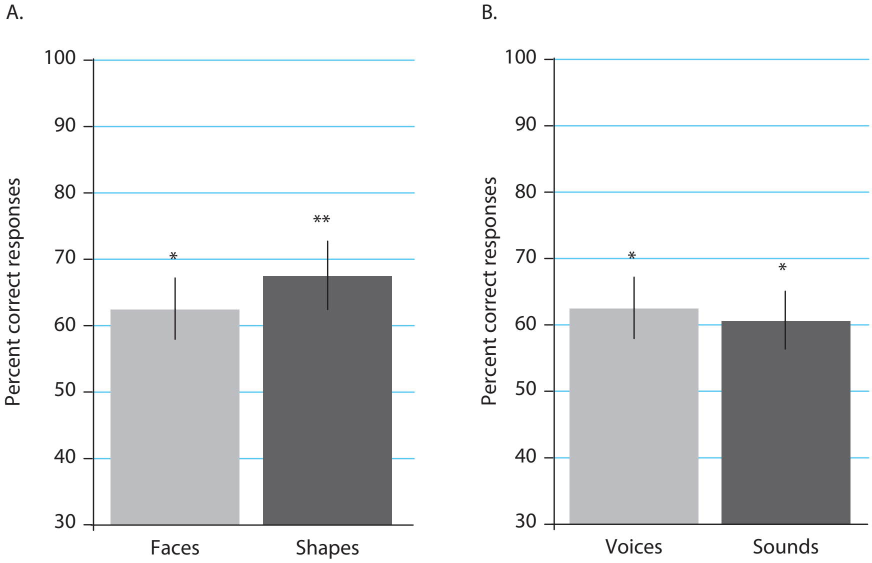

**Figure 2.** Learning in Experiment 1, sequential grouping of unimodal visual and auditory signals. **A.** Learning of visual stimuli. The y-axis shows the percentage of correct responses, with the line at 50% showing random performance. The error bars show ± standard error of the mean. Single asterisk denotes significance at $p < 0.05$ and double asterisk denotes significance at $p < 0.01$. **B.** Learning of auditory stimuli. The y-axis shows the percentage of correct responses, with the line at 50% showing random performance. The error bars show ± standard error of the mean. Single asterisk denotes significance at $p < 0.05$

## 3. Experiment 2: Visual elements coupled with auditory elements

A particular feature of faces and voices is that, in the social environment, they are coupled: the face of one individual is associated with a specific voice. This redundancy across modalities can enhance recognition, discrimination and learning of individuals. In this experiment, we tested whether subjects could implicitly learn sequential groups of faces in parallel with groups of voices when each face was reliably coupled with one specific voice. We compared this learning to coupled shapes and syn-sounds. As in Experiment 1, the stimulus sets were divided into short sequences of three elements, base-triplets. The visual and auditory element sequences were presented in parallel with each other. Each face (or shape) was coupled with one specific voice (or syn-sound) (Figure 3A).

We predicted that the coupling of visual elements with specific auditory elements should enhance statistical learning, as it does during other types of multisensory learning (Lehmann & Murray 2005; Seitz, Kim, & Shams 2006; Shams & Seitz 2008), including paired-associates learning (Seitz, Kim, Van Wassenhove, & Shams 2008). As attention modulates the statistical learning of one set of visual

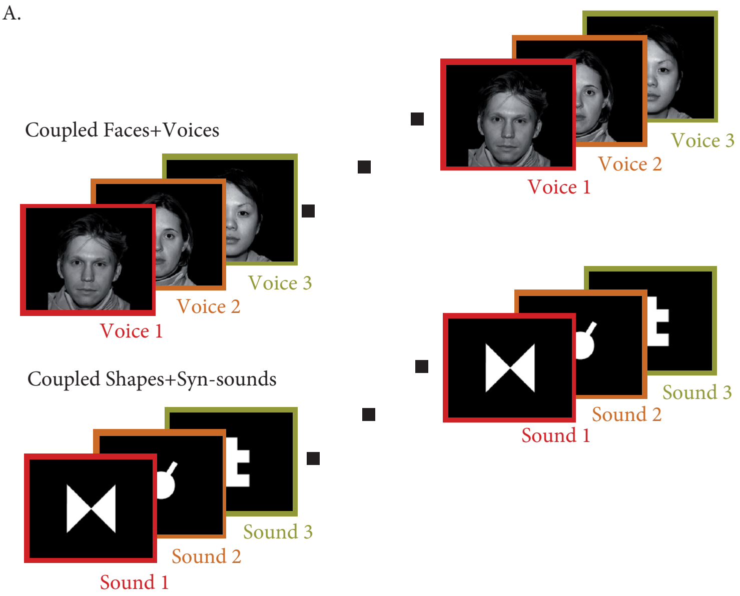

**Figure 3.** Illustration of the familiarization-phase stimulus structures for experiments 2 to 4. **A.** Experiment 2, visual elements coupled with auditory elements. Note that faces are presented together with voices and shapes together with syn-sounds, and that the same visual stimulus is always presented together with the same auditory stimulus. *(Continued)*

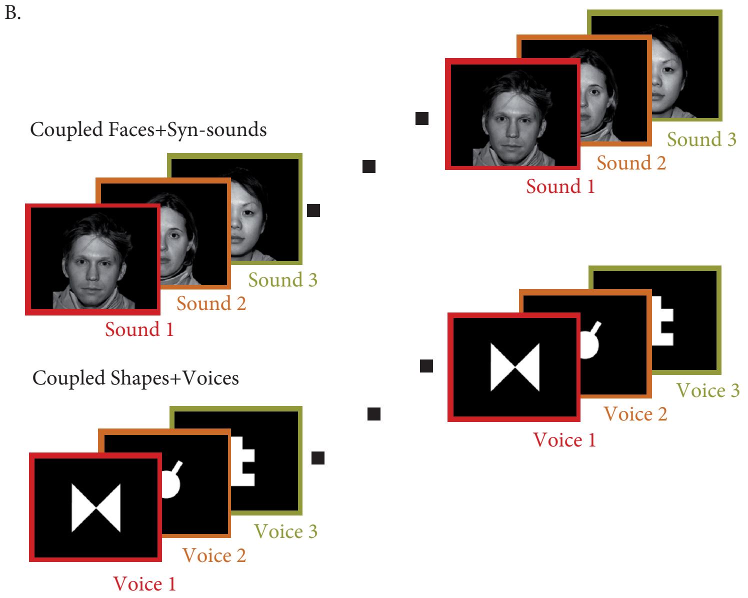

**Figure 3. B.** Experiment 3, faces coupled with syn-sounds and shapes coupled with voices. Note faces are presented with syn-sounds and shapes with voices, and the same visual stimulus is always presented together with the same auditory stimulus. (Continued)

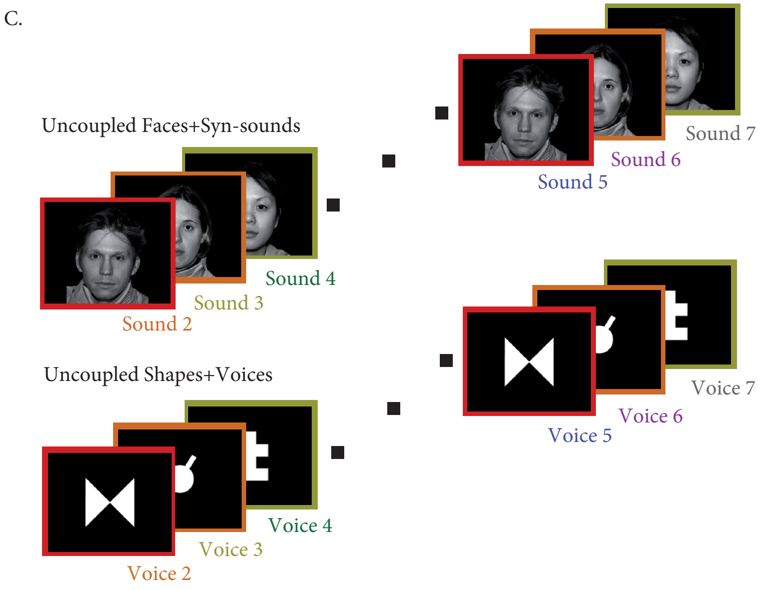

**Figure 3. C.** Experiment 4, faces uncoupled from syn-sounds and shapes uncoupled from voices. Note faces are presented with syn-sounds and shapes with voices, whereas the same visual stimulus is not always presented together with the same auditory stimulus. (Continued)

shapes versus another shown in the same familiarization sequence (Turk-Browne et al. 2005), an alternative hypothesis is that the coupling of elements across modalities could serve to distract attention away from one stream versus the other and thus disrupt learning. In addition, we expected that the face/voice coupling should lead to better overall statistical learning than shape/syn-sound coupling, although the learning of the visual-auditory association on an individual basis may be the same for both stimulus classes (von Kriegstein & Giraud 2006).

### 3.1 Methods

*Subjects.* Twenty-two naïve subjects participated in this experiment. Subjects participated in two counterbalanced conditions: one with faces and voices and the other with shapes and syn-sounds.

*Stimuli.* The visual and auditory elements were the same as in Experiment 1.

*Design.* We used two simultaneously running familiarization streams (one visual and one auditory) generated in the same way as in Experiment 1, with the difference that they now were audio-visual. The visual and auditory streams were coupled, so that each individual visual element always co-occurred with the same individual auditory element (Figure 3A).

*Procedure.* The procedure was similar to Experiment 1, with the difference that subjects were now tested for visual, auditory, audio-visual base-triplets. They were subsequently tested for their recognition of audio-visual pairs — the element wise association between visual and auditory elements. During the familiarization phase, subjects were exposed to 2 × 96 base-triplets for a total period of approximately 2 minutes and 50 seconds. During the test phase, a base-triplet and a cross-triplet were played sequentially, separated by 1 second. Cross-triplets were generated as in Experiment 1. Subjects were tested with visual only, auditory only, and audiovisual base-triplets: each modality consisted of 8 test trials, with a total of 24 trials. Each pair of base-triplets and cross-triplets was shown twice with their order counterbalanced. For every subject, the order of test triplets was randomized (i.e. unimodal and bimodal tests were intermingled).

To investigate if subjects learned the coupling (association) between individual visual and auditory elements (von Kriegstein & Giraud 2006), we also tested subjects on the 12 different audio-visual pairs (outside of the 'triplet' context). After finishing the familiarization phase and the test of base-triplets, they were presented with a series of audio-visual trials that consisted of one matching audio-visual pair that had been presented to them during the familiarization phase and another non-matching pair that did not occur together during the familiarization phase. Every matching audio-visual pair was tested once. Since subjects were exposed to the same number of audiovisual pairs in both the base-triplets and the cross-triplets, the only opportunity for them to learn the pairs was during the familiarization phase. Thus, the prior testing of base-triplets could not influence their performance. The order of match and non-match audio-visual pairs was randomized and counterbalanced across the 12 trials.

### 3.2 Results and discussion

Statistical learning of coupled face/voice triplets was similar to coupled shape/syn-sound triplets (Figure 4A). We performed a 2 × 3 × 2 mixed ANOVA with stimulus class (human and synthetic) and stimulus modality (visual, auditory and visual-auditory) as within subject factors, and condition order as a betweenubjects factor. There was no significant main effect for stimulus class ($F(1, 20) = 0.49$, $p = 0.49$), a significant main effect of modality, with performance on visual and visual-auditory stimuli being better than auditory alone ($F(2,40) = 4.48$, $p = 0.018$), no interaction between class and modality ($F(2,40) = 3.66$, $p = 0.035$), and finally, no significant effect of condition order ($F(1, 20) = 1.25$, $p = 0.277$). All conditions were learned better than chance, as assessed by a one-sample t-test (Figure 4A and Table 1).

Falsifying our prediction, there was no enhancement of recognition via the bimodal presentation of elements (Figure 4A and Table 1). One explanation for this is that, given the short (under 3 minutes) familiarization phase consisting of a somewhat rapid sequence of elements, the subjects simply did not detect that the elements from the visual sequence were coupled with those in the auditory sequence. To test this, we tested subjects on their ability to detect familiar versus unfamiliar visual-auditory pairs. For both face/voice pairs and shape/syn-sound pairs, subjects performed significantly above chance (face/voice, $t(21) = 10.17$, $p << 0.001$; shape/syn-sound, $t(21) = 7.33$, $p << 0.001$) (Figure 4B). Thus, the subjects readily learned that certain visual elements belong with certain auditory elements (see also, von Kriegstein & Giraud 2006).

In summary, this experiment shows that when auditory and visual elements are coupled, they were learned in parallel. Faces coupled with voices did not have a special status in terms of statistical learning when compared to the arbitrarily related shapes and synthetic sounds. Furthermore, subjects readily learned that the particular face or shape elements were coupled to specific voice or syn-sound elements (Figure 4B). Finally, there was no enhanced recognition of triplets when they were presented crossmodally versus unimodally in the test condition. However, the significant main effect of stimulus modality suggests that performance on auditory triplets was worse than performance on visual or visual-auditory triplets. This impaired performance appears especially to be true for synthetic sounds.

Table 1. Experiment 2

| | T | Df | P (2-tailed) |
|---|---|---|---|
| Faces | 5.109 | 21 | <<0.001 |
| Voices | 3.634 | 21 | 0.002 |
| Faces + Voices | 4.297 | 21 | <<0.001 |
| Shapes | 5.667 | 21 | <<0.001 |
| Syn-sounds | 2.570 | 21 | 0.018 |
| Shapes + Syn-sounds | 7.339 | 21 | <<0.001 |

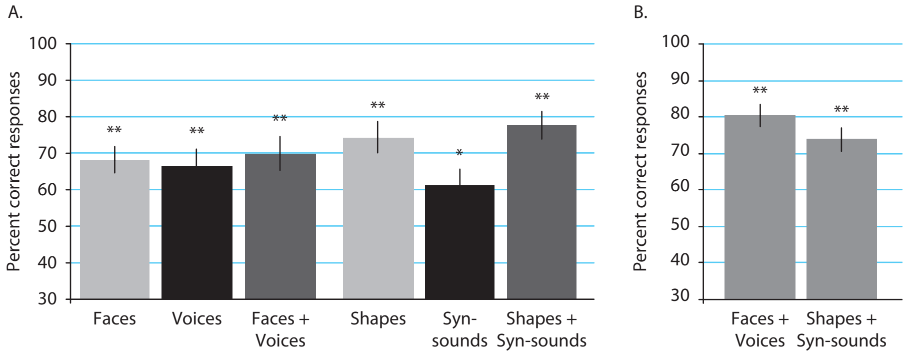

**Figure 4.** Learning in experiment 2, visual elements coupled with auditory elements. **A.** Left half: Learning of faces, voices, and face-voice combinations. Right half: Learning of syn-sounds, shapes, and syn-sound-shape combinations. The *y*-axis shows the percentage of correct responses, with the dashed line at 50% showing chance performance. The error bars show ± the standard error of the mean. Single asterisk denotes significance at $p < 0.05$, double asterisk denotes significance at $p < 0.01$. **B.** Learning of audio-visual pairs. The *y*-axis shows the percentage of correct responses, with the dashed line at 50% showing chance performance. The error bars show ± the standard error of the mean. Double asterisk denotes significance at $p < 0.01$

## 4. Experiment 3: Faces coupled with syn-sounds and shapes coupled with voices

In this experiment, we investigated whether statistical learning of sequences was different when faces were coupled with syn-sounds and voices were coupled with shapes (Figure 3B). This tests whether same or different processes mediate the parallel learning of faces with voices or shapes with syn-sounds. If it is the same process, then coupling faces and syn-sounds should lead to parallel learning, as should coupling shapes with voices. If social signals are learned through different processes than artificial signals, then parallel learning should be absent.

### 4.1 Methods

The stimuli, design and procedure of this experiment are identical to Experiment 2, except that faces were coupled with syn-sounds and voices were coupled with shapes (Figure 3B). Twenty-six subjects participated in this experiment, counterbalanced according to condition.

### 4.2 Results and discussion

In Experiment 3, subjects were presented with faces coupled with syn-sounds and voices coupled with shapes (Figure 3B). Remarkably, their performance was nearly identical to that seen in Experiment 2. That is, parallel learning was observed for the two modalities, regardless of which class of stimuli were paired together. We performed a 2 × 3 × 2 mixed ANOVA with stimulus class (human, synthetic) and stimulus modality (visual, auditory, visual-auditory) as within-subject factors, and condition order as a between-subjects factor. We did not find significant main effects for stimulus class ($F(1, 24) = 0.65$, $p = 0.43$), stimulus modality ($F(2, 48) = 0.04$, $p = 0.96$), or an interaction between class and modality ($F(2,48) = 1.47$, $p = 0.24$), nor did we find a significant effect of condition order ($F(1, 24) = 0.03$, $p = 0.86$). As in the other experiments, the latter suggesting subjects' participation in two consecutive experiments did not influence learning.

Even though a main effect of stimulus modality is lacking, the outcome of this experiment is comparable to Experiment 2. As in Experiment 2, performance on syn-sounds tends to be lower than on corresponding visual triplet, however in this case; syn-sounds were combined with faces, and not shapes. In contrast, for voices the opposite effect is observed, performance is better than on shapes, suggesting that modality is not the only factor influencing performance. Thus, the lack of a main effect of modality suggests that both stimulus class and modality influences learning.

In essence, the results show that parallel learning was evident for all conditions (Figure 5A and Table 2). When faces were coupled with syn-sounds, triplet sequences in both modalities were learned. The same was true for the coupling of shapes with voices. As in Experiment 2, there was no evidence for enhanced recognition of triplet sequences when they were presented as bimodal signals (faces/syn-sounds or voices/shapes) in the test phase. This lack of enhancement cannot be attributed to a failure to learn that particular visual elements were associated with particular auditory elements (Figure 5B). Subjects learned that faces were associated with specific syn-sounds ($t(25) = 5.73$, $p << 0.001$) and that shapes were associated with specific voices ($t(25) = 5.37$, $p << 0.001$). These data suggest that, in the domain of statistical learning, faces and voices are independently permissive for the parallel learning of any signals from the other modality that are coupled to them.

To test the possibility that the shapes and syn-sounds are indeed the signals that are "permitted", we uncoupled the visual and auditory streams in the next experiment.

**Table 2.** Experiment 3

| | T | Df | P (2-tailed) |
|---|---|---|---|
| Faces | 3.005 | 25 | 0.006 |
| Syn-sounds | 2.409 | 25 | 0.024 |
| Faces + Syn-sounds | 2.900 | 25 | 0.008 |
| Shapes | 2.570 | 25 | 0.008 |
| Voices | 4.698 | 25 | <<0.001 |
| Shapes + Voices | 2.570 | 25 | 0.017 |

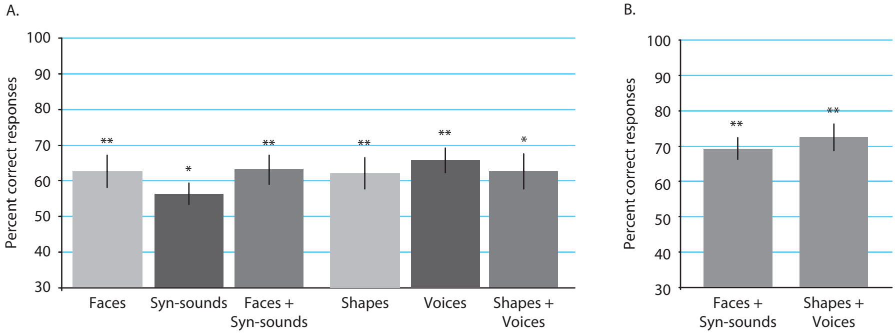

**Figure 5.** Learning in experiment 3, faces coupled with syn-sounds and shapes coupled with voices. **A.** Left half: Learning of faces, syn-sounds, and face-syn-sound combinations. Right half: Learning of shapes, voices, and shape-voice combinations. The *y*-axis shows the percentage of correct responses, with the dashed line at 50% showing chance performance. The error bars show ± the standard error of the mean. **B.** Learning of audio-visual pairs. The *y*-axis shows the percentage of correct responses, with the dashed line at 50% showing chance performance. The error bars show ± the standard error of the mean. Double asterisk denotes significance at $p < 0.01$

## 5. Experiment 4: Faces uncoupled from syn-sounds and shapes uncoupled from voices

When faces and voices are coupled with arbitrary elements from another modality, parallel learning is evident across all conditions (Experiment 3). This suggests faces and voices could be acting as permissive signals for the learning of those arbitrary elements that are coupled to them. We tested whether this was true by presenting faces in parallel with, but uncoupled from, syn-sounds, and the same for voices with shapes (Figure 3C). That is, although visual and auditory streams were presented in parallel, each visual element was no longer paired with a specific auditory element. This tests whether the parallel learning seen in Experiments 2 and 3 is due to the special salience of faces and voices and the *coupling* of elements across modalities. If faces and voices were special, then we would expect to see only learning of faces, but not syn-sounds in one condition, and only learning of voices, but not shapes in the other condition.

### 5.1 Methods

The stimuli, design and procedure of this experiment is identical to Experiment 3, except that faces were presented in parallel with uncoupled syn-sounds, and voices were presented in parallel with uncoupled shapes (Figure 3C). Eighteen naïve subjects participated in this experiment; conditions were counterbalanced.

### 5.2 Results and discussion

This experiment revealed that, when uncoupled, faces and voices dominated over syn-sounds and shapes during statistical learning (Figure 3C). We performed a 2 × 2 × 2 mixed ANOVA with stimulus class (human, synthetic) and stimulus modality (visual, auditory) as within-subject factors, and condition order as between-subjects factor. We did not find any significant main effect of stimulus class ($F(1, 16) = 2.32$, $p = 0.15$), or a main effect for stimulus modality ($F(2, 16) = 0.17$, $p = 0.68$), but we did see a significant interaction between stimulus class and stimulus modality ($F(2, 16) = 7.31$, $p < 0.05$). As in all the other experiments, there was no significant effect of condition order, but there was a trend ($F(1, 16) = 3.82$, $p = 0.07$). However, since subjects were counterbalanced across condition order, such a trend does not confound the results.

To explore the significant interaction between stimulus class and modality, we conducted a series of two tailed t-tests comparing the results to chance. When faces were uncoupled from syn-sounds, faces were learned (70.1%; $t(17) = 3.70$, $p < 0.01$, but syn-sounds were not (56.3%, $t(17) = 1.64$, $p = 0.12$). Similarly, when voices were uncoupled from shapes, voices were learned (63.3%; $t(17) = 3.70$, $p < 0.01$), but shapes were not (52.1%; $t(17) = 0.51$, $p = 0.61$) (Figure 6). Taken together with the results of Experiment 3, these data suggest that faces and voices are independently permissive to learning arbitrary signals from another modality but only when they are coupled on an individual element-by-element basis. To put it another way, when social signals are conflicted with arbitrary signals, the social signals have priority.

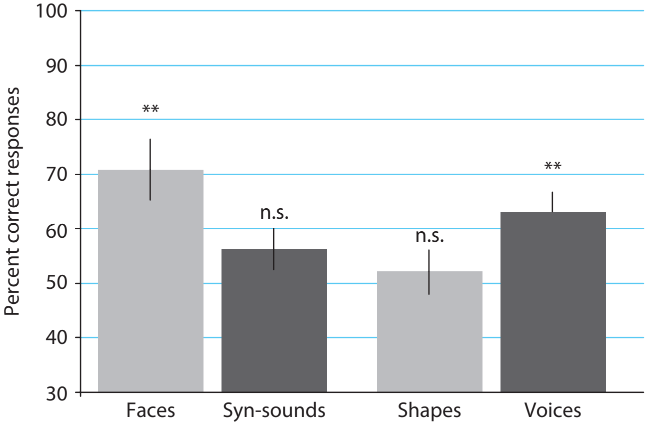

**Figure 6.** Learning in experiment 4, faces uncoupled from syn-sounds and shapes uncoupled from voices. Left half: Learning of faces, and syn-sounds. Right half: Learning of shapes, and voices. The *y*-axis shows the percentage of correct responses, with the dashed line at 50% showing chance performance. The error bars show ± the standard error of the mean. Double asterisk denotes significance at $p < 0.01$ and n.s. denotes $p > 0.05$

## 6. General discussion

We were interested in mechanisms by which primates could learn their social group structure without specialized cognitive modules. We tested whether human subjects could learn triplets of social signals — faces and voices — in a statistical learning paradigm and, if so, whether this learning was better or worse when compared to arbitrary signals. From the outset, there were four possible outcomes to our study: (1) Since faces and voices are among the most salient features of the human environment, statistical learning could be better for these signals than they would be for artificial stimuli; (2) Because faces and voices seem to be processed by specialized perceptual and neural strategies, the statistical learning paradigm may not operate, or operate sub-optimally, for social signals; (3) Faces and voices and artificial shapes and sounds are learned equally well under a statistical learning paradigm, suggesting a domain-general mechanism; and (4) The pre-existing crossmodal association of faces and voices leads to better statistical learning when compared to arbitrary associations of objects and sounds. This latter possibility differs from the first outcome in that a difference between faces and voices, on the one hand, and artificial stimuli, on the other, could become apparent only with cross-modally presented stimuli because faces and voices have a natural association familiar to humans that arbitrary shapes and sounds do not.

We found that, when presented unimodally, faces and voices were learned just as well, but not better, than arbitrary shapes and sounds (Experiment 1). When faces and voices were coupled on an element-by-element basis, they were learned in parallel, but again coupled arbitrary signals were learned just as well (Experiment 2). This was somewhat surprising because individual faces are typically associated with individual voices, while no such relationship exists for arbitrary shapes and sounds. To test this further, we coupled faces to arbitrary sounds (syn-sounds), and shapes to voices (Experiment 3). Here again parallel learning of triplets in both modalities was evident, supporting the existence of a domain general statistical learning mechanism. It appears that, while faces and voices may be special in other domains of behavior, they are not special for statistical learning. However, uncoupling the social signals from the arbitrary ones (whereby visual and auditory streams were still presented in concurrently, but in which each visual element was not associated with a specific auditory element) resulted only in the learning of faces and voices (Experiment 4). This suggests that when social signals are put in conflict with arbitrary signals presented in a different modality, then social signals have perceptual priority.

### 6.1 Domain generality and parallel learning across modalities

Our data show that, when presented unimodally, faces and voices are statistically learned with the same robustness as shapes and syn-sounds. This is somewhat surprising. Given that faces and voices are such special and salient signals for humans, processed by specialized brain networks, we predicted that triplets of such signals would either be learned *better* than less ecologically-relevant signals or that they would be *not* be learned as well, perhaps because we automatically look at faces and hear voices as individuals not as groups. Thus, our data strongly support the idea that statistical learning is a domain general mechanism — learning is irrespective of the elements that instantiate the pattern, at least when presented unimodally. This is an important extension to previous findings, as most studies investigating visual statistical learning used abstract stimuli, with little ecological relevance, such as abstract shapes, colors, or spatiotemporal patterns (but see Saffran and colleagues who used images of cats and dogs (J. R. Saffran, Pollak, Seibel, & Shkolnik 2007)). Similarly, in auditory statistical learning, tones and synthesized speech are often used and the synthesized speech lacks the indexical cues that allow the listener to extract information about the speaker's identity and physical characteristics (size, age, gender, etc)(Ghazanfar & Rendall 2008).

Differences in learning between the two stimulus classes — social and arbitrary — also did not appear when the two modalities were presented concurrently. We had two general predictions. First, we reasoned that, regardless of signal category, if individual visual elements were reliably paired with auditory elements, then learning should be enhanced, as is the case for other types of multisensory learning (Shams & Seitz 2008). For example, explicit training with congruent audio-visual stimuli improves performance in visual only motion detection compared to training with visual only stimuli (Seitz et al. 2006). Similarly, visual image recognition is enhanced when past experience of it included a semantically congruent sound, but is impaired when paired with an incongruent sound (Lehmann & Murray 2005). Surprisingly, when subjects were tested for recognition using bimodal cues, we found no multisensory enhancement of recognition for either stimulus category, suggesting that not all types of learning benefit from multisensory experience (Shams & Seitz 2008).

Second, we predicted that, since faces are naturally coupled to voices in the real world, while arbitrary shapes and syn-sounds are not, coupled faces and voices should be learned better than coupled shapes and syn-sounds. This was not the case. We found that faces and voices could be learned in parallel when they were coupled, as were shapes coupled with synthetic sounds. These data, too, suggest a domain general mechanism for statistical learning, but it extends the mechanism to multisensory statistical learning. Importantly, in the coupled conditions, subjects readily (and implicitly) learned that individual visual elements were associated with specific auditory elements (see also von Kriegstein & Giraud 2006 for a similar result). This contingency between individual elements across modalities was important. We found that coupling faces with syn-sounds or voices with shapes resulted in parallel statistical learning of all types of elements, but uncoupling the same signals resulted in only statistical learning of faces (but not syn-sounds) and voices (but not shapes). Thus, coupling elements to either faces or voices has a permissive effect for parallel statistical learning, but when they are uncoupled (i.e. put into conflict), social signals have perceptual priority. In the statistical learning context, faces and voices appear to be strong signals that can 'carry' whatever is associated with them.

This begs the question as to why faces and voices are prioritized. The unimodal data suggest that the statistical learning mechanism seems to be the same for faces, voices, shapes and syn-sounds. The multisensory data suggest that as long as there is element-to-element coupling, then any category of stimuli paired with faces or with voices will be learned in parallel. If the two streams are uncoupled, the faces and voices will take priority over the arbitrary element sequence. Together, this finding suggests one of two possibilities: (1) faces and voices engage more sensory resources than do other categories of stimuli (though not necessarily specialized resources); and/or (2) faces and voices engage more efficient mechanisms of sensory processing. The dedication of more resources is likely linked to the overwhelming experience with such social signals that humans acquire pre- and post-natally (Leppanen & Nelson 2008; Lewkowicz & Ghazanfar 2009).

### 6.2 Implications for the social brain hypothesis

Humans may engage each other in a highly action centered, continuous, spatial jockeying for position and influence within the confines of the group (at work, at play, in the dormitory, etc), using social contact and proximity as a means to achieving immediate goals, and monitoring the action of others. There are two mechanisms by which they may learn about this dynamic social structure. In the representational framework, social knowledge would be mediated through specialized modules that process only one kind of information and are distinct from other modules. For example, the recognition of social signals such as faces and voices that seem to be processed by specialized brain areas in the temporal lobe supports this notion (Belin, et al. 2004; McKone, et al. 2007). In contrast, in an active perception framework, social signals and knowledge are acquired through generic pattern recognition (Barrett, et al. 2007). Instead of specialized modules, social signals and structure are embedded in the patterns of activation of neuronal units, linked in distributed neural networks. Thus, faces, for example, are processed by multiple areas simultaneously, in regions that are not dedicated solely to faces, but are activated to a greater spatial extent and with a greater magnitude by faces than other visual signals (Haxby, et al. 2001). Furthermore, such networks are unlikely to be purely unimodal — further evidence against any strict modular organization can be found in (Ghazanfar & Schroeder 2006).

Our data suggest the statistical learning mechanism, an implicit form of learning, can be used to learn about social group structures using a distributed associative network. For long-lived primates, like humans, the physical and social environments are both inherently unstable. Given these constrains, statistical learning provides a mechanism arguably more adaptive than specialized cognitive routines, allowing fast (less than 3 minutes are needed to track three individuals), flexible and experientially informed pattern recognition to form the basis for much of social cognition. In fact, humans can identify coalitional alliances based on attire and spatiotemporal associations in less than 4 minutes of exposure (Kurzban, et al. 2001). However, as Barrett et al. (2007) have argued, the potential downside of such a generic pattern recognition mechanism is that it is tissue intensive: large-scale pattern recognizers require a lot of connectivity to implement (Clark 1993). Thus, according to Barrett et al.'s hypothesis (and consistent with our results), the social brain does not get bigger with increasing group size due to the adding of specialized modules, but rather the relationship between group size and neocortical size rises from the more generic requirement of a larger associative network able to deal with bigger social groups.

## Acknowledgments

We thank Stephen Shepherd and Chand Chandrasekaran for their helpful comments on previous versions of this manuscript. This work was supported by a National Science Foundation BCS-0547760 CAREER Award (AAG).

## References

Barrett, L., & Henzi, P. (2005). The social nature of primate cognition. *Proceedings of the Royal Society B, 272*, 1865–1875.

Barrett, L., Henzi, P., & Rendall, D. (2007). Social brains, simple minds: does social complexity really require cognitive complexity? *Philosophical Transactions of the Royal Society B, 362*, 561–575.

Barrett, L., & Rendall, D. (2009). Out of our minds: the neuroethology of primate strategic behaviour. In M. L. Platt & A. A. Ghazanfar (Eds.), *Primate neuroethology* (pp. In press). Oxford, UK: Oxford University Press.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129–135.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305.

Cheney, D.L., & Seyfarth, R.M. (2007). *Baboon metaphysics: the evolution of a social mind.* Chicago: The University of Chicago Press.

Clark, A. (1993). *Associative engines: connectionism, concepts, and representational change.* Cambridge, MA: MIT Press.

Clark, A. (1997). *Being there: putting brain, body, and world together again.* Cambridge, MA: MIT Press.

Conway, C.M., & Christiansen, M.H. (2005). Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 24–39.

Conway, C.M., & Christiansen, M.H. (2006). Statistical learning within and between modalities: pitting abstract against stimulus-specific representations. *Psychol Sci, 17*(10), 905–912.

Dunbar, R.I.M. (1992). Neocortex Size as a Constraint on Group-Size in Primates. *Journal of Human Evolution, 22*(6), 469–493.

Dunbar, R.I.M. (1995). Neocortex Size and Group-Size in Primates – a Test of the Hypothesis. *Journal of Human Evolution, 28*(3), 287–296.

Dunbar, R.I.M. (1998). The social brain hypothesis. *Evolutionary Anthropology, 6*(5), 178–190.

Dunbar, R.I. M., Duncan, N.D. C., & Nettle, D. (1995). Size and Structure of Freely Forming Conversational Groups. *Human Nature: an Interdisciplinary Biosocial Perspective, 6*(1), 67–78.

Eisenberg, J.F. (1965). The social organisation of mammals. *Handbook of Zoology, 8*, 1–92.

Fiser, J., & Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol Sci, 12*(6), 499–504.

Fiser, J., & Aslin, R.N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *J Exp Psychol Learn Mem Cogn, 28*(3), 458–467.

Fiser, J., & Aslin, R.N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America, 99*(24), 15822–15826.

Ghazanfar, A.A., & Rendall, D. (2008). Evolution of human vocal production. *Curr Biol, 18*, R457-R460.

Ghazanfar, A.A., & Santos, L.R. (2004). Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience, 5*(8), 603–616.

Ghazanfar, A.A., & Schroeder, C.E. (2006). Is the neocortex essentially multisensory? *Trends Cogn Sci, 10*, 278–285.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425–2430.

Howard, J.H., Howard, D.V., Dennis, N.A., & Kelly, A.J. (2008). Implicit learning of predictive relationships in three-element visual sequences by young and old adults. *Journal of Experimental Psychology-Learning Memory and Cognition, 34*(5), 1139–1157.

Humphrey, N. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology.* (pp. 303–317). Cambridge, MA: Cambridge University Press.

Johnson, C.M. (2001). Distributed primate cognition: a review. *Animal Cognition, 4*, 167–183.

Jolly, A. (1966). Lemur social behavior and primate intelligence. *Science, 153*, 501–506.

Kudo, H., & Dunbar, R.I.M. (2001). Neocortex size and social network size in primates. *Animal Behaviour, 62*, 711–722.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 15387–15392.

Lehmann, S., & Murray, M.M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research, 24*(2), 326–334.

Leppanen, J.M., & Nelson, C.A. (2008). Tuning the developing brain to social signals of emotions. *Nature Reviews Neuroscience, 10*, 37–47.

Lewkowicz, D.J., & Ghazanfar, A.A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends Cogn Sci, 13*, 470–478.

McKone, E., Kanwisher, N., & Duchaine, B.C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11*(1), 8–15.

Perez-Barberia, F.J., Shultz, S., & Dunbar, R.I.M. (2007). Evidence for coevolution of sociality and relative brain size in three orders of mammals. *Evolution, 61*(12), 2811–2821.

Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence.* Cambridge, MA: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical Learning by 8-Month-Old Infants. *274*(5294), 1926–1928.

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

Saffran, J.R., Pollak, S.D., Seibel, R.L., & Shkolnik, A. (2007). Dog is a dog is a dog: infant rule learning is not specific to language. *Cognition, 105*(3), 669–680.

Scherf, K.S., Behrmann, M., Humphreys, K., & Luna, B. (2007). Visual category-selectivity for faces, places and objects emerges along different developmental trajectories. *Dev Sci, 10*(4), F15-30.

Seitz, A.R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Curr Biol, 16*(14), 1422–1427.

Seitz, A.R., Kim, R., Van Wassenhove, V., & Shams, L. (2008). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception, 36*, 1445–1453.

Shams, L., & Seitz, A.R. (2008). Benefits of multisensory learning. *Trends Cogn Sci, 12*, 411–417.

Sugita, Y. (2008). Innate face processing. *Current Opinion In Neurobiology, 19*, 39–44.

Tsao, D.Y., Cadieu, C.F., & Livingstone, M.S. (2010). Object recognition: physiological and computational insights. In M. L. Platt & A. A. Ghazanfar (Eds.), *Primate neuroethology* (pp. 471–499). Oxford: Oxford University Press.

Turk-Browne, N.B., Jung, J., & Scholl, B.J. (2005). The Automaticity of Visual Statistical Learning. *Journal of Experimental Psychology: General, 134*, 552–564.

von Kriegstein, K., & Giraud, A.L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biol, 4*(10), e326.

Werker, J.F., & Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 7*(1), 49–63.

## *Authors' addresses*

Hjalmar K. Turesson
Department of Psychology
Green Hall
Princeton University
Princeton NJ 08540
USA

Email: turesson@princeton.edu

Asif A. Ghazanfar (corresponding author)
Neuroscience Institute
Department of Psychology
Green Hall
Princeton University
Princeton NJ 08540
USA

Email: asifg@princeton.edu

## *Biographical notes*

**Mr. Hjalmar Turesson** is a Ph.D. candidate in the Department of Psychology at Princeton University. His thesis work focuses on the structure of sensory signals, their influence on vocal communication and how such an influence is mediated by neocortical circuits. He received his B.Sci. in Biology from Lund University in Sweden.

**Dr. Asif Ghazanfar** is an Associate Professor in the Neuroscience Institute and Departments of Psychology and Ecology & Evolutionary Biology at Princeton University. His research focuses on the neurobiology and evolution of primate communication and how both aspects are influenced by body morphology, development and socioecological context. Asif received his B.Sci. in Philosophy from the University of Idaho and his Ph.D. in Neurobiology from Duke University. He was a postdoctoral fellow at Harvard University and a research scientist at the Max Planck Institute for Biological Cybernetics in Tübingen, Germany before moving to Princeton.