Primate ACC encodes natural vocal interactions in a 'cocktail party'

Arthur Lefevre[1,2], Vikram Pal Singh[1], Timothy Tyree[1], Jingwen Li[1], Jean-René Duhamel[2] & Cory Miller[1]

1. Cortical Systems & Behavior Laboratory, University of California San Diego, USA

2. Institute of Cognitive Sciences Marc Jeannerod, CNRS, University Lyon 1, France

Summary

The Cocktail Party Problem (CPP)—extracting meaningful signals amid competing voices—remains poorly understood at the neural level, particularly in real-world contexts where it emerges naturally. We investigated the role of the anterior cingulate cortex (ACC) to resolve the CPP, a structure implicated in social monitoring but rarely examined in relation to audition, in freely-moving marmoset monkeys engaged in vocal exchanges with conspecifics in a noisy ecological environment. Analyses revealed that this neural substrate is seemingly integral to resolving the CPP. Not only did neurons encode the calls of either the conversational partner or background callers (i.e. figure/ground separation), but this selectivity persisted even with overlapping background sounds, a hallmark of the CPP. ACC was also sensitive to conversational dynamics by encoding turn-taking structure. These findings reveal that the ACC regulates dynamic vocal interactions in complex social soundscapes in which parsing meaningful calls from competing sounds is essential for effective communication.

Main Text

The CPP is arguably the most persistent real-world challenge for acoustic communication systems, such as speech[1]. Little remains known about the underlying neural mechanisms because, first, most neurobiological research related to the CPP has predominantly been performed in controlled laboratory settings, emphasizing the sensory aspects of CPP, rather than in the acoustically dynamic, naturalistic environments where the CPP was first described[2] and naturally arises. Second, studies have focused almost exclusively on neocortical substrates involved in human speech processing, such as secondary fields of auditory cortex[3–5] and ventrolateral prefrontal cortex[6], while mostly overlooking other neural structures involved in social monitoring and cognition. This bias is particularly notable given that the CPP predates both human speech and the emergence of these neocortical structures in primate evolution. The challenges for social communication in natural contexts are not solely acoustic; these social interactions rely on cognitive processes and related neural mechanisms beyond traditional speech and language circuits[7]. As such, elucidating how the brain resolves the CPP may require considering a broader range of neural substrates. One such candidate is the anterior cingulate cortex (ACC), which has been largely ignored in studies of CPP despite its established role in vocal production[8] and more recently vocal perception[9], and, critically, social monitoring during interactions with conspecifics[10].

We hypothesized that the ACC will exhibit at least the three following mechanisms integral to resolving the CPP in natural communication scenes: (i) selectivity for call type and social context, (ii) respond invariantly to vocalizations regardless of background sounds and (iii) encode structure of active vocal turn taking exchanges.

Because resolving the CPP enables individuals to communicate with conspecifics in complex natural environments[11,12], we posited that uncovering its neural basis likely requires its study under similarly dynamic conditions[13]. To this end, we leveraged the rich vocal communication system of marmoset monkeys[14] — highly vocal New World primates[15,16]. Marmosets engage in a range of context-

dependent vocal exchanges, including turn taking with different call types[12], and produce vocalizations that are modulated by social dynamics[17–19].

Segregated ACC neuronal populations encode vocalizations perceived from the partner and the colony

To examine the neural basis of the CPP in marmosets, all experiments were performed in the noisy, colony room. The subject (2 females, 1 male) was placed in a small transparent box in front of their home cage (Fig. 1a). We recorded ACC area 24 wirelessly using micro wire brush arrays or Neuropixels probes (Fig. 1b, Extended Data Fig. 1), yielding 1599 neurons and 38559 vocalizations across 46 sessions (Fig. 1b, Extended Data Fig. 2). Using an automated labeling algorithm[20], analyses distinguished whether each call was produced by the subject, their partner, or the colony.

Similar to previous work[8,21–24], we observed that ACC neurons were predominantly motor modulated, though a population were responsive to hearing the calls of conspecifics (i.e vocalization perception) (Fig. 1c). While ACC exhibits sensory responses to passive presentations of vocalizations[9,25], whether this reflects a general response to conspecific calls or neurons in this substrate will display a selectivity to different call types necessary to resolve the CPP during natural vocal communication contexts was not clear. To test this, we first analyzed ACC neurons activity in response to hearing their partner's vocalizations and those produced by other conspecifics in the colony (Fig. 1d). Analyses revealed that neurons in area 24 responded robustly to both partner calls (Wilcoxon signed rank tests, $p < 0.05$, $n = 167$, Fig. 1e, Extended Data Fig. 3c) and to colony calls (Wilcoxon signed rank tests, $p < 0.05$, $n = 77$, Fig. 1f). The majority of neurons responded to only a single call type (86.8%, Extended Data Fig. 3d). Given this degree of call type specificity in the activity of individual ACC neurons, we investigated whether call types were robustly encoded at the population level. A multiclass decoder revealed that individual call types from both the partner (Wilcoxon signed rank test, $z = 4.16$, $p < 0.05$) (Fig. 1g, Extended Data Fig. 4) and the colony (Wilcoxon signed rank test, $z = 3.19$, $p < 0.05$) (Fig. 1h, Extended Data Fig. 4) could be significantly decoded from ACC population activity.

We next examined whether ACC area 24 neurons separately encode different types of callers in the scene, namely partner calls and those produced by other marmosets in the colony. In other words, do these neurons simply encode call type or a combination of call type and social categories? We observed two distinct populations of neurons, one that responded only when hearing partner calls (Figure 1i) and the second that responded only to hearing other callers in the colony (Figure 1j). This suggests that ACC neurons parse conspecific calls into two categories - partner and others (i.e. colony) – and encode both. While these neural populations were almost entirely segregated; 2 neurons (out of 244) responded to both partner and colony calls, suggesting the existence of a minority of neurons with general auditory responses. Moreover, the selectivity was not biased to partner calls, with similar proportions of neurons encoding both categories of callers within the same ACC substrate (5.3% of neurons encoding partner phees or twitters, 4.8% encoding colony phees or twitters). At the very least, these neurons suggest that figure/ground separation of highly meaningful signals[1,2,26] is encoded in ACC area 24, a foundational mechanism for resolving the CPP in natural acoustic scenes. To further validate this idea at the population level, we used neural population activity to decode whether a call had been produced by the partner or the colony using a support vector machine (SVM) classifier. Analyses revealed significant decoding of these social categories (Wilcoxon signed rank test, $z = 2.91$, $p < 0.05$) (Fig. 1k).

ACC neurons respond invariantly to calls with and without overlapping vocalizations

To fully resolve the CPP, neural mechanisms that parse meaningful signals from acoustically overlapping background sounds are necessary[1,2], particularly in the presence of similar vocalizations produced by conspecifics. To more closely examine how ACC responses were affected by background sounds, we investigated how neurons responded to vocalizations in the two following contexts[26]: 'Simultaneous calls' (i.e. calls that occurred concurrently with a colony call (Fig. 2a)) and '"Sequential calls" (i.e. calls that had no acoustic overlap with other vocalizations (Fig. 2b)[26]). Analyses focused on

twitter calls because it had the largest population of perceptual neurons selective to partner calls (n = 54 neurons), thereby allowing comparison in both contexts.

Analyses revealed that although neurons were strongly driven to partner twitter calls (Wilcoxon signed rank tests, $p < 0.05$), there was no statistical difference in responses between the sequential and simultaneous contexts ($p > 0.05$) (Fig. 2c, d), indicating that background sounds do not affect the responses of these neurons. In other words, single neurons in marmoset ACC area 24 selectively parse the meaningful signal – vocalizations produced by partners - from overlapping background sounds (e.g., all other conspecific vocalizations), thereby exhibiting a key mechanism necessary to resolve the CPP. Following this logic, populations of neurons encoding partner calls and colony calls were almost entirely distinct (Fig. 1j, k). This suggests that attentional mechanisms in ACC[27–29] enable neurons in this substrate to encode distinct calls produced by conspecifics that are temporally simultaneous.

Marmosets Monitor Social Scenes to Optimize Vocal Turn Taking in a Cocktail Party

Resolving the CPP is essential for vocal communication in both human and nonhuman animals, where the ability to parse, produce, and respond to vocal signals amidst background noise enables meaningful interactions. To better understand how marmosets communicate in such environments, we examined vocal turn taking with different vocalizations in the noisy colony environment, a natural behavior observed in many animals including humans and marmosets[12].

To quantify marmoset turn-taking in this contexts, we focused analyses on two distinct vocalization types — trills[14,20] and phees[14,31] (Fig. 3a, b) — by aligning the partner and colony call rates on the subject's vocalizations (Fig. 3c, d). Trills are low-amplitude calls exchanged between affiliated conspecifics in close proximity[14,20], whereas phees are higher-amplitude, long-distance contact calls, with less affiliative component[14,31]. This natural ethological difference between these call types during turn-taking provided an opportunity to investigate the behavioral strategies marmosets use to

communicate in noisy acoustic scenes. We found that partners produced significantly more trills in close temporal proximity to the subject's trill onset (sliding Wilcoxon's signed rank tests, -0.2 to 1s, $p < 0.05$, Fig. 3c) while they avoided pheeing concomitantly (0 to 1.4s, $p < 0.05$, Fig. 3c). Moreover, subjects produced the low amplitude trill calls when the colony phee call rate was significantly lower than average (-2 to -0.2s, $p < 0.05$, Fig. 3d), effectively avoiding noisy periods in the colony, and phee calls when the colony phee call rate was higher than average (-2.2 to 0.1s, $p < 0.05$, Fig. 3d), presumably to respond to other individuals. This pattern suggests that marmosets monitor the colony's sound level and adopt context-dependent strategies for turn-taking with each call type to maximize respective communication signaling efficacy.

ACC neurons encode active turn taking with the partner and the colony

Although the CPP has been considered mainly as a perceptual problem, it is integral to communication in noisy socio-acoustic environments. Thus, in the present paradigm, our subjects were not only perceiving calls but also producing them, reflecting active communication. We analyzed neural activity during vocal production. Results showed that neurons exhibited a significant increased ($n = 311$, 19.4%) or decreased ($n = 652$, 40.8%) activity when the subject was producing a call (Wilcoxon signed rank tests, $p < 0.05$) (Fig. 4a-c, Extended Data Fig. 3a, b, d). Like call perception (Fig. 1f, g), the majority of neurons were highly selective to only one call type (83.3%, Extended Data Fig. 3d). However, a subset of neurons exhibited responses to multiple types of vocalizations (Extended Data Fig. 3d).

We quantified whether produced call type could be robustly decoded from neural population activity using a multiclass decoder trained on calls produced by subjects. Results showed that call types could be accurately classified based on ACC activity (Wilcoxon signed rank test, $z = 4.54$, $p < 0.05$) (Fig. 4d, Extended Data Fig. 5). This decoding was dependent on the number of neurons rather than on the ones significantly modulated by call production (Extended Data Fig. 5). Together, these data show that

ACC area 24 neurons encode call production with a high degree of selectivity, consistent with results in other motor systems[30].

To investigate whether ACC neurons are sensitive to the social context in which calls are produced, we employed a time sensitive decoding analysis designed to identify periods with maximal differences in neural activity between different contexts (see Methods and [32] for details). Specifically, we applied this approach to test whether population activity could reliably classify whether a given call occurred as part of a structured turn-taking exchange or independent of these social interactions. Turn-taking was defined as an event in which subjects produced a call within 5 seconds (for partner Trill calls) or 10 seconds (for colony Phee calls) of a conspecific producing the same call type (Fig. 3a, b).

Analyses revealed that the behavioral context – whether the vocalization occurred in a turn-taking exchange or not – could be accurately classified from ACC area 24 population activity. Trill calls produced (accuracy: 86.9%) or perceived (accuracy: 82.7%) (Fig. 4e, Extended Data Fig. S6c) were accurately classified as isolated or part of a turn-taking bout with the partner, according to one-sided Mann-Whitney's U-tests ($z = 2.48\times10^5$, $p < 0.0001$, $N_{calls} = 596$ produced, $N_{calls} = 776$ perceived). Likewise, phee calls produced (accuracy: 74.0%) or perceived (accuracy: 97.6%) (Fig. 4f, Extended Data Fig. S6d) were classified with significant statistical accuracy according to a one-sided Mann-Whitney's U-test (produced: $z = 8.2\times10^4$, $p < 0.0001$, $N_{calls} = 304$; perceived: $z = 8.8\times10^4$, $p < 0.0001$, $N_{calls} = 381$).

These results indicate that ACC area 24 encodes not only the calls themselves, but information about social context. Because many of these turn-taking exchanges occurred during noisy acoustic periods, particularly for phee calls, this also suggests that levels of background noise do not affect ACC activity during a coordinated social interaction, further emphasizing the role of this neural substrate to resolve the CPP.

Discussion

Here, we investigated the role of ACC area 24 in resolving the CPP as freely-moving marmosets engaged in natural vocal communication in a noisy environment. We found that single neurons selectively responded to partner vocalizations or colony calls, effectively parsing different categories of conspecific speakers in the acoustic scene (Fig. 1). Critically, neurons selective for partner calls responded invariantly when calls overlapped acoustically with other sounds, showing that neural mechanisms to distinguish meaningful voices from other interfering voices, are present in marmoset ACC area 24 (Fig. 2). These findings suggests that ACC may be a keystone neural substrate for resolving the CPP, highlighting the significance of this process for mediating social interactions in natural scenes, rather than solely a challenge for audition.

As a midline forebrain area conserved across all mammals[33] that is both anatomically connected to fronto-lateral[34] and auditory cortex[35], and modulated by attention[27–29], it may perhaps not be surprising that the ACC supports resolution of the CPP. Although not classically considered part of the language network, results here show that ACC neurons are involved in a range of communicative functions, including representing each of the call types produced or perceived in vocal exchanges (Fig. 1, 3). Recent marmoset fMRI studies have indeed implicated ACC in vocalization processing[9,25], suggesting strong interactions with the broader communication network. Moreover, if we view the CPP as a non-literal language comprehension problem, resolving it requires listeners to recognize each speaker's voice. Afferent connections from hippocampus, which robustly encodes identity in social signals[32], to area 24[35], could provide this necessary input. Taken together, these findings and the underlying anatomical connectivity suggest that ACC area 24 plays a more central role in resolving the CPP than previously thought, potentially emphasizing that the CPP is not merely a challenge for audition but one of sociality.

Marmosets engage in vocal turn-taking[12,14,36,37]. In the noisy colony setting, we observed differences in the relative timing of the call types produced in these social exchanges in the presence of background

noise. Whereas high amplitude, long-distance phee calls were produced in synchrony with other phee calls from the colony, trill calls – a call used when in close visual contact - were mostly produced during periods of silence in the colony. This shows that marmosets monitor the acoustic landscape to optimize signaling efficacy during these communicative exchanges[38–41]. Moreover, ACC appears to play a role in this process, as the social context in which these calls were produced and perceived could be reliably decoded from neural activity alone (Fig. 4). In parallel, fMRI findings in humans point out that ACC is activated by listening to conversations rather than monologues[42], and more generally that ACC key role in theory of mind is integral to natural language processing[43–46], and language production, as ACC lesions in humans produce akinetic mutism[47]. Finally, ACC functions for social monitoring[48] and exploration[49] make it a logical node to be recruited for processing the CPP.

By capturing neural activity during natural vocal exchanges of freely-moving primates in rich acoustic environments, we discovered that the anterior cingulate cortex area 24—a region long overlooked in this domain—implements core computations necessary for resolving the CPP. Certainly, ACC is not solely responsible for resolving the CPP, but results here do suggest a likely central role of this neural substrate for integrating the requisite processes necessary for this foundational auditory process. Moreover, these results emphasize that CPP computations extend beyond sound parsing to orchestrating social communication in noise by integrating perception, action, and context, highlighting the role of this area in mediating social communication. This work showcases the transformative potential of naturalistic neuroscience and points to a broader reevaluation of how and where the brain solves real-world cognitive challenges[13].

References

1. McDermott, J. H. The cocktail party problem. *Curr. Biol. CB* **19**, R1024-1027 (2009).

2. Bee, M. A. & Micheyl, C. The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? *J. Comp. Psychol. Wash. DC 1983* **122**, 235–251 (2008).
3. Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).
4. O'Sullivan, J. *et al.* Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron* **104**, 1195-1209.e3 (2019).
5. Nelken, I., Bizley, J., Shamma, S. A. & Wang, X. Auditory cortical processing in real-world listening: the auditory system going real. *J. Neurosci. Off. J. Soc. Neurosci.* **34**, 15135–15138 (2014).
6. Joshi, N. *et al.* Temporal coherence shapes cortical responses to speech mixtures in a ferret cocktail party. *Commun. Biol.* **7**, 1392 (2024).
7. Klein, J. T., Shepherd, S. V. & Platt, M. L. Social attention and the brain. *Curr. Biol. CB* **19**, R958-962 (2009).
8. Jürgens, U. Neural pathways underlying vocal control. *Neurosci. Biobehav. Rev.* **26**, 235–258 (2002).
9. Jafari, A. *et al.* A vocalization-processing network in marmosets. *Cell Rep.* **42**, 112526 (2023).
10. Putnam, P. T. & Chang, S. W. C. Social processing by the primate medial frontal cortex. *Int. Rev. Neurobiol.* **158**, 213–248 (2021).
11. Nieder, A. & Mooney, R. The neurobiology of innate, volitional and learned vocalizations in mammals and birds. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **375**, 20190054 (2020).
12. Burkart, J. M. *et al.* A convergent interaction engine: vocal communication among marmoset monkeys. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **377**, 20210098 (2022).
13. Miller, C. T. *et al.* Natural behavior is the language of the brain. *Curr. Biol. CB* **32**, R482–R493 (2022).
14. Bezerra, B. M. & Souto, A. Structure and Usage of the Vocal Repertoire of Callithrix jacchus. *Int. J. Primatol.* **29**, 671–701 (2008).

15. Digby, L. J. & Barreto, C. E. Social organization in a wild population of Callithrix jacchus. I. Group composition and dynamics. *Folia Primatol. Int. J. Primatol.* **61**, 123–134 (1993).
16. Lazaro-Perea, C. Intergroup interactions in wild common marmosets, *Callithrix jacchus*: territorial defence and assessment of neighbours. *Anim. Behav.* **62**, 11–21 (2001).
17. Miller, C. T. *et al.* Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**, 219–233 (2016).
18. Burkart, J. M. & van Schaik, C. P. Marmoset prosociality is intentional. *Anim. Cogn.* **23**, 581–594 (2020).
19. Vitale, A., Zanzoni, M., Queyras, A. & Chiarotti, F. Degree of social contact affects the emission of food calls in the common marmoset (Callithrix jacchus). *Am. J. Primatol.* **59**, 21–28 (2003).
20. Landman, R. *et al.* Close-range vocal interaction in the common marmoset (Callithrix jacchus). *PloS One* **15**, e0227392 (2020).
21. Robinson, B. W. Vocalization evoked from forebrain in Macaca mulatta. *Physiol. Behav.* **2**, 345–354 (1967).
22. Sperli, F., Spinelli, L., Pollo, C. & Seeck, M. Contralateral smile and laughter, but no mirth, induced by electrical stimulation of the cingulate cortex. *Epilepsia* **47**, 440–443 (2006).
23. Gavrilov, N., Hage, S. R. & Nieder, A. Functional Specialization of the Primate Frontal Lobe during Cognitive Control of Vocalizations. *Cell Rep.* **21**, 2393–2406 (2017).
24. West, R. A. & Larson, C. R. Neurons of the anterior mesial cortex related to faciovocal activity in the awake monkey. *J. Neurophysiol.* **74**, 1856–1869 (1995).
25. Dureux, A., Zanini, A., Trapeau, R., Belin, P. & Everling, S. Functional organization of voice patches in marmosets and cross-species comparisons with macaques and humans. *Curr. Biol. CB* S0960-9822(25)00874–7 (2025) doi:10.1016/j.cub.2025.07.008.
26. Bregman, A. S. Auditory scene analysis: Hearing in complex environments. (1993).
27. Petersen, S. E. & Posner, M. I. The attention system of the human brain: 20 years after. *Annu. Rev. Neurosci.* **35**, 73–89 (2012).

28. Schneider, K. N., Sciarillo, X. A., Nudelman, J. L., Cheer, J. F. & Roesch, M. R. Anterior Cingulate Cortex Signals Attention in a Social Paradigm that Manipulates Reward and Shock. *Curr. Biol. CB* **30**, 3724-3735.e2 (2020).
29. Benedict, R. H. B. *et al.* Covert auditory attention generates activation in the rostral/dorsal anterior cingulate cortex. *J. Cogn. Neurosci.* **14**, 637–645 (2002).
30. Li, J., Aoi, M. C. & Miller, C. T. Representing the dynamics of natural marmoset vocal behaviors in frontal cortex. *Neuron* **112**, 3542-3550.e3 (2024).
31. Miller, C. T., Mandel, K. & Wang, X. The communicative content of the common marmoset phee call during antiphonal calling. *Am. J. Primatol.* **72**, 974–980 (2010).
32. Tyree, T. J., Metke, M. & Miller, C. T. Cross-modal representation of identity in the primate hippocampus. *Science* **382**, 417–423 (2023).
33. Burgos-Robles, A., Gothard, K. M., Monfils, M. H., Morozov, A. & Vicentic, A. Conserved features of anterior cingulate networks support observational learning across species. *Neurosci. Biobehav. Rev.* **107**, 215–228 (2019).
34. Ducret, M. *et al.* Medial to lateral frontal functional connectivity mapping reveals the organization of cingulate cortex. *Cereb. Cortex N. Y. N 1991* **34**, bhae322 (2024).
35. Vogt, B. A. & Pandya, D. N. Cingulate cortex of the rhesus monkey: II. Cortical afferents. *J. Comp. Neurol.* **262**, 271–289 (1987).
36. Bosshard, A. B. *et al.* Beyond bigrams: call sequencing in the common marmoset (Callithrix jacchus) vocal system. *R. Soc. Open Sci.* **11**, 240218 (2024).
37. Oren, G. *et al.* Vocal labeling of others by nonhuman primates. *Science* **385**, 996–1003 (2024).
38. Eliades, S. J. & Wang, X. Neural correlates of the lombard effect in primate auditory cortex. *J. Neurosci. Off. J. Soc. Neurosci.* **32**, 10737–10748 (2012).
39. Tsunada, J. & Eliades, S. J. Frontal-auditory cortical interactions and sensory prediction during vocal production in marmoset monkeys. *Curr. Biol. CB* S0960-9822(25)00393–8 (2025) doi:10.1016/j.cub.2025.03.077.

40. Löschner, J., Pomberger, T. & Hage, S. R. Marmoset monkeys use different avoidance strategies to cope with ambient noise during vocal behavior. *iScience* **26**, 106219 (2023).

41. Löschner, J. & Hage, S. R. Sound amongst the din: primate strategies against noise. *Trends Cogn. Sci.* **29**, 111–113 (2025).

42. Olson, H. A., Chen, E. M., Lydic, K. O. & Saxe, R. R. Left-Hemisphere Cortical Language Regions Respond Equally to Observed Dialogue and Monologue. *Neurobiol. Lang. Camb. Mass* **4**, 575–610 (2023).

43. Fedorenko, E., Ivanova, A. A. & Regev, T. I. The language network as a natural kind within the broader landscape of the human brain. *Nat. Rev. Neurosci.* **25**, 289–312 (2024).

44. Ferstl, E. C. & von Cramon, D. Y. What does the frontomedian cortex contribute to language processing: coherence or theory of mind? *NeuroImage* **17**, 1599–1612 (2002).

45. Amodio, D. M. & Frith, C. D. Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* **7**, 268–277 (2006).

46. Wittmann, M. K., Lockwood, P. L. & Rushworth, M. F. S. Neural Mechanisms of Social Cognition in Primates. *Annu. Rev. Neurosci.* **41**, 99–118 (2018).

47. Barris, R. W. & Schuman, H. R. [Bilateral anterior cingulate gyrus lesions; syndrome of the anterior cingulate gyri]. *Neurology* **3**, 44–52 (1953).

48. Clairis, N. & Lopez-Persem, A. Debates on the dorsomedial prefrontal/dorsal anterior cingulate cortex: insights for future research. *Brain J. Neurol.* **146**, 4826–4844 (2023).

49. Kolling, N., Behrens, T. E., Mars, R. B. & Rushworth, M. F. Neural Mechanisms of Foraging. *Science* **336**, 95 (2012).

50. Paxinos, G., Watson, C., Petrides, M., Rosa, M. & Tokuno, H. *The Marmoset Brain in Stereotaxic Coordinates*. (Elsevier Academic Press, 2012).

51. McMahon, D. B. T., Bondar, I. V., Afuwape, O. A. T., Ide, D. C. & Leopold, D. A. One month in the life of a neuron: longitudinal single-unit electrophysiology in the monkey visual system. *J. Neurophysiol.* **112**, 1748–1762 (2014).

52. Oikarinen, T. *et al.* Deep convolutional network for animal sound classification and source attribution using dual audio recordings. *J. Acoust. Soc. Am.* **145**, 654 (2019).
53. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & D, H. K. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. 061481 Preprint at https://doi.org/10.1101/061481 (2016).
54. Tyree, T. J. Applications of Mathematical Physics to Quantitative Biology. (2023).
55. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in 785–794 (2016). doi:10.1145/2939672.2939785.
56. Jovanovic, V., Fishbein, A. R., de la Mothe, L., Lee, K.-F. & Miller, C. T. Behavioral context affects social signal representations within single primate prefrontal cortex neurons. *Neuron* S0896-6273(22)00059–9 (2022) doi:10.1016/j.neuron.2022.01.020.

Figure legends

Figure 1. ***Segregated ACC neuronal populations encode vocalizations perceived from the partner and the colony.*** **a)** Schema of the experimental paradigm. *Top* – recordings took place directly in the home cage and transfer box in the colony room. *Center* - both subject and partner monkeys were equipped with a portable microphone, and the implanted subject was also equipped with a neurologger. *Bottom* – Neurons were recorded with brush arrays or Neuropixel probes implanted in ACC area 24. **b)** Coronal plane (AP = +12mm) at which probes were implanted. **c)** Percentage of neurons responding to vocalization production and perception. **d)** Vocalizations produced by the partner or other marmosets in the colony (vs produced ones) were analyzed separately. **e)** Normalized average increase from neurons with significantly upregulated activity during partner call perception. **f)** Normalized average increase from neurons with significantly upregulated activity during colony call perception. **g-h)** Percentage accuracy of the neuronal multiclass decoder for partner **g)** and colony **(h)** perceived call type for real data and data with shuffled labels. **i)** Example neuron responding specifically to twitter calls from the partner but not from the colony. **j)** Example neuron responding specifically to phee calls

from the colony but not from the partner. **k)** Percentage accuracy of the neuronal SVM decoder for identity (partner vs colony) for real data and data with shuffled labels. *: $p < 0.05$, error bars indicate s.e.m, (standard error to the mean).

Figure 2. ***ACC neurons respond invariantly to calls with and without overlapping vocalizations.*** **a)** Example spectrograms from the colony and partner microphones showing the occurrence of a call from the partner (twitter) that is perceived simultaneously with a call from the colony (phee). **b)** Example spectrograms from the colony and partner microphones showing the occurrence of a call from the partner (twitter) that is perceived sequentially to calls from the colony (twitters). **c)** Example neurons responding to twitter calls from the partner that are unaffected by the presence (or absence) of simultaneous calls from the colony. **d)** Normalized average increase from neurons with significantly upregulated activity during partner twitter call perception is unaffected by the presence (or absence) of simultaneous calls from the colony. *: $p < 0.05$, error bars indicate s.e.m.

Figure 3. ***Marmosets Monitor Social Scenes to Optimize Vocal Turn Taking in a Cocktail Party.*** **a)** Example spectrograms showing trill to trill conversations between the subject and its partner. **b)** Example spectrograms showing phee to phee conversations between the subject and other monkeys in the colony. **c)** Normalized partner call rate for trills (purple) and phees (orange) aligned on subject's trills and phees onsets. **d)** Colony call rate for phees aligned on subject's trills (purple) and phees (orange).

Figure 4. ***ACC neurons encode active turn taking with the partner and the colony.*** **a)** Normalized average increase from neurons with significantly upregulated activity during call production. **b)** Example neuron responding to the production of several call types. **c)** Normalized average decrease

from neurons with significantly downregulated activity during call production. **d)** Percentage accuracy of the neuronal multiclass decoder for produced call type for real data and data with shuffled labels. **e-f)** Receiver Operator Characteristic (ROC) traces for our neural decoder trained on predictive time bins representing whether trills answered trills with the subject's partner in panel **e**, the answering of phee calls with the colony in panel **f**. Blue solid lines represent calls initially produced by the subject while orange solid lines represent calls initially perceived by the subject. Random chance is represented by the red dashed line as a guide. The area under the curve (AUC) quantifies the success of the neural decoder as our outcome measure. *: $p < 0.05$, error bars indicate s.e.m.

Methods

Animals

Four adult marmosets (*Callithrix jacchus*) living in bonded pairs (1 female and 1 male per cage) for at least 3 months were used for this experiment. All pairs were housed in a room with other marmoset cages including two adjacent cages with visual separators (between 41 and 47 marmosets in total). Three of the marmosets, monkey S (female), G (female) and L (male) were implanted with electrodes while monkey K (male) was only used for behavior. Animals had unrestricted access to food and water. All experiments were approved by the UCSD Institutional Animal Care and Use Committee and were performed in the Cortical Systems and Behavior Laboratory at University of California San Diego (UCSD).

Behavioral paradigm

Marmosets were first habituated to be handled and taken out in transport boxes by the experimenters for a month. Once habituated to handling (measured by assessing the comfort of animals while being handled and taking treats from the experimenters' hands), we gradually habituated them over one or two months to wear the leather harness containing the microphone until they spent less than 20% of the time chewing on it. Experimental sessions lasted between 1 hour to 1.5 hours. We recorded only

one pair of monkeys per session. During a session, the pair was taken out of their home cage in transport boxes and brought to an adjacent experimental room for preparation. We then equipped both the subject and the partner monkey with the custom-made leather harness containing a microphone (micro voice recorder, Spycentre Security®). Additionally, we connected a neurologger to the subject's head-cap. Both monkeys were brought back to the colony room. The partner was placed back in its home-cage and the subject was kept in a transparent transport box placed at the entrance of the home-cage (Fig 1a, Fig S2a) giving the pair visual and auditory access to each other. To record colony calls, we used an omnidirectional microphone (H3-VR, Zoom®) placed in the middle of the room. This paradigm allowed us to record acoustic interactions between the subject and its partner, and between the subject and the colony. In marmosets, which are capable of turn-taking, these interactions involve the coordinated exchange of vocal signals between individuals, characterized by social coordination and minimal acoustic overlap, allowing speakers to alternate the timing of their calls within a shared acoustic space.

Neurophysiological procedures

To target the desired brain regions for electrophysiological recording using external probes, the brain anatomy of the subjects was imaged using a 3T MRI (Siemens MAGNETOM Prisma) scanner and a human knee coil (Siemens/QED Transmit/Receive 15 channels). T2 and T1 weighted images of the whole marmoset brain were obtained in an anaesthetized animal placed in a custom non-magnetic 3D printed stereotax with earbars and bite plate. We adapted the coordinates from the Paxinos atlas[50] of the marmoset brain based on individual brain size. We targeted the left hemisphere area 24 (Site 1: AP +12mm, ML +1mm, Site 2: AP +13.5mm, ML +1mm). On the day of surgery, monkeys were intubated and maintained under isoflurane anesthesia, then placed in a surgical stereotaxic frame (Kopf). We performed a skin incision and a craniotomy (1.5mm diameter), then inserted a ground electrode between the dura and the skull before covering the skull with metabond dental cement (C&B

Metabond®). We employed 2 different strategies. For monkey S and G, we used micro wire brush arrays (Microprobe Maryland, USA, Monkey S:1x64 channels, Monkey G: 2x32 channels with 2mm space in between) with microdrives (custom made from [51], for Monkey G, an additional hole was pierced and the drive adapted to hold 2 micro brush arrays). The electrode guide tube(s) (Microlumen, High performance medical tubing, 310-I.5 PTFE), containing a 24G needle, attached to the microdrive was/were inserted straight into the brain at a depth of 1.5mm. The needle(s) was/were retracted and the electrode(s) array(s) were lowered into the guide tube(s) until they protruded out of the guide tube by 1mm. Kwik Sil (World Precision Instrument) was used to fill the craniotomy and seal the gap at the top of the guide tube(s). Then the Microdrive and the omnetics connector were cemented to the metabond using dental acrylic, after soldering the ground electrode to the array ground wire. Finally, the skin was glued to the metabond using Vetbond (3M). For monkey L, we employed Spikegadget wireless neuropixel system. The procedure was similar, except that a durotomy was performed and the 2 probes (Npx 1.0 and Npx NHP 1.0 10mm) inserted 5mm deep from the cortical surface before we covered them with kiwk-Sil and fixed them directly to the skull with metabond. Probes were then connected to the Spikegadget head cap that was built around the electrodes and cemented to the metabond. After the surgery, monkeys received analgesic treatment for 3 days. Recording sessions started 2 weeks after the surgery for monkey S and G and were performed twice per week, and started 4 days after surgery and happened every day for monkey L (monkey S: 21 sessions, monkey G: 19 sessions, Monkey L: 6 sessions). Implantation sites were verified by structural MRI for monkey G and S (Fig S1b), but this could not be done for monkey L.

To wirelessly record neurons, we used a 64 channels Neurologger (Deuteron Technologies) for monkey S and G with a sample rate of 32 KHz, and Neuropixels Datalogger Headstage (SpikeGadgets) with a sample rate of 30 KHz for monkey L.

Data preprocessing

Vocalizations: To automatically label subjects' and partners' vocalizations, we retrained the algorithm described in [20] on the data recorded from our 4 monkeys with the wearable microphones. This neural network[52] identifies the call type and determine the source (subject or partner) by comparing the amplitude. If the amplitude is equal in both microphones, the call is considered to be coming from the colony. Following this step, all audio files were manually corrected to remove eventual mislabeling by the algorithm and determine precisely (<20ms) each call onset and offset. We considered only 4 call types: Trill, Twitter, Phee and Chirp. Other calls (Tsik, Ek, Chatter, TrillPhee, Whistle) had too few occurrences to be analyzed. Additionally, to investigate the cocktail party effect, Twitter calls coming from the partner were labeled as "simultaneous" or "sequential" depending on if a call from the colony was overlapping or not with it.

For calls coming from the colony (recorded with the omnidirectional microphone), a first pass was done with ACDC neural network (https://github.com/mineraldragon/ACDC_2022) and then corrected manually. We only labeled Twitter and Phee calls from the colony, as other call types such as Trill and Chirp were not loud enough to be recorded consistently, or had too few occurrences of them (eg Tsik, Chatter, etc.). Note that for technical reasons, colony calls were not always recorded (34 out of 46 sessions with colony recordings).

We removed all calls that were happening less than 2 seconds after another call from the same source, to ensure that neuronal responses analyzed were not influenced by a recent prior call. This was the case for all analyses except for PSTH of phee calls perceived from the partner (Fig 1f), due to low number of calls. Thus, Fig 1f phee call data may contain residual activity from successive calls.

Finally, we removed calls that occurred when the neural signal was poor due to artifacts, defined as calls for which 20% of the neural signal (from 1sec before to 1 sec after the call onset) was labeled as artifact (see neural data).

Neural data: All neural data were band pass filtered (300-7000Hz). Signal from brush arrays (but not Neuropixel) contained artifacts, that were removed by setting signal value to 0 if the standard

deviation of one channel was above a threshold (arbitrarily set to 100 after visual inspection of signal). If more than 20% of a session was removed, the session was not included in our analyses. We then used Kilosort 2[53] to performed automatically the spike sorting for both micro brush arrays and Neuropixels data. Each identified unit was manually checked and validated and labeled as single or multi-unit or rejected based on the waveform shape and auto-correlogram using Phy (open-source Python library for spike visualization and curation). Only units with a minimum firing rate of 0.5Hz were considered. Across the 3 monkeys (Monkey S: units, Monkey G: 574 units, Monkey L: 794 units) a total of 1599 units were identified, of which 932 were well isolated single units (using the PCA). However, because we did not see differences between single and multi-units in later analyses, we decided to pull them together and to refer to them (single and multi-neurons) as neurons throughout this article. All PSTH in this article are showing results from well isolated single units.

Synchronization: All data streams (3 microphones and neurologger) were synchronized using 250ms audio tone pulses played via a buzzer generated by an ESP32 microcontroller programmed using Arduino IDE.

All .wave files were manually aligned with Audacity® (v 3.4). Two tones (one at the start of the session and another at the end) were used to bookend the whole session which allowed us to identify and correct the temporal drift of each microphone in Audacity using the “change speed” functionality. Audacity was able to correct the temporal drift in each microphone to less than 20ms. Following this, a MATLAB script was used to randomly remove data points until the temporal time drift was reduced to <1ms. This did not influence the labeling of calls by the artificial neural network.

A 100 mv TTL pulse (100 ms) was sent to the Neurologger using the same ESP32 microcontroller neuron at the beginning of the session. The time difference between the Neurologger TTL and first acoustic pulse was used to align the timing of vocalizations onset and offset to the neurologger time. Using this approach, we were able to synchronize different systems with varying sampling frequencies to a common master clock.

Statistical analyses

PSTH responses: Firing rate was computed over 100ms bins. To test neurons' response to produced or perceived calls we first employed a peak detection approach. For each call type, if there were at least 5 calls, we calculated the mean firing rate between -5s to -2s before call onset for produced calls and between -5sec to 0sec before call onset for perceived calls as a baseline. We then searched for the peak response between -2s and +1s around call onset for produced calls and between 0sec and 3sec after call onset for calls perceived. To ensure a minimum peak width, we then tested if the bins around the peak bin reached at least 50% of the peak value (threshold = baseline +- 0.5*(peak-baseline)) to have at least 3 consecutive bins (300ms) beyond this threshold. Finally, we tested if the peak values were significantly higher than the baseline values using the non-parametric paired Wilcoxon signed rank test. We used this approach to allow for time flexibility because we observed that neurons responded at various time windows (Fig S4, 6). Note that both maximum and minimum peak responses were computed to look for excited and inhibited neurons. Each neuron was tested for each produced and perceived call types.

Finally, the demonstrational PSTH traces reported in Figure S7 were first computed at a bin width of 0.3 seconds, then up-sampled to a bin width of 0.1 seconds using linear interpolation that assigned knots to the mean time of any given bin, and finally smoothed with a sliding window with a duration of 2 seconds using a Savitzky-Golay filtration. Uncertainty of these demonstrational PSTH traces indicates 95% confidence estimated via bootstrap and is shown by the relatively horizontal, yet curved shaded regions in Figure S7.

Decoding: To decode call type or caller category, we employed a multiclass decoder (error-correcting output codes model, MATLAB function *fitcecoc*) with a one-versus-all coding design. For each session, the input was the average firing rate of each neuron between -2s and +1s around call onset for produced calls and between 0sec and 3sec after call onset for calls perceived. The number of calls of each type (Trill, Twitter, Phee or Chirp) or caller (partner or colony) was equalized by subsampling

categories with more calls down to the category that had the least. Only calls that had at least 10 occurrences were considered, leading to different number of categories and therefore chance levels for each session (25%, 33% or 50%). The decoder was trained on 80% of the data and tested on the remaining 20%. To control for random effects, we also decoded on neural data with shuffled labels. This procedure was repeated 100 times to obtain a representative average decoding value for the data and the shuffle. Finally, we used Wilcoxon sign rank test to compare the average decoding accuracies obtained on the real and the shuffled data.

Time bins: To quantify the neural representations that identified whether a Trill with the subject's partner was answered or whether a Phee with the subject's colony was answered, we used a predictive time bin analysis similar to previous work in the lab[32,54]. In this study, we applied the same four-fold stratified cross-validation both to detect predictive time bins and to decode their apparent firing rates using an ensemble of gradient-boosted decision trees[55]. Recording sessions were selected for these analyses if they had at least ten repeatedly spiking neurons recorded, at least four recorded calls that were answered, and at least four recorded calls that were not answered. A Trill call was considered an answer if it happened within 5 seconds of a Trill call from the other monkey, and a Phee call was considered an answer if it happened within 10 seconds of a Phee call from other colony monkeys. These timings were based on previous literature [20,56]. As a control, the same analysis was repeated with pseudo-randomly shuffled labels, and decoding results are reported in the main text and in Figure 4. All time bins considered only the times from five seconds before to five seconds after call onset and were constrained to be no briefer than four hundred milliseconds in duration, resulting in a median activation time of t=-0.17 (IQR: -2.34 through +2.14) seconds (N_{bins} = 4326 predictive time bins) relative to call onset at t=0. Predictive time bins were also constrained to be non-overlapping for any given neuron such that no action potential was considered twice for a given type of neural representation. Reported tree-based decoder accuracies were computed using a threshold parameter value that maximized apparent accuracy. Area under the ROC curves provide an alternative outcome measure that is parameterless and is reported in the Fig 4d-g.

We found 2139 significant predictive time bins across 948 neurons (Extended Data Fig. 7a, c) that were significantly predictive if a Trill call was part of a turn-taking bout in terms of median firing rate (two-sided Mann-Whitney's U-test, $p < 0.05$). Only 1098 neurons were considered in these Trill call analyses by the minimal session selection criteria, which was used equally in all predictive time bin analyses. We found 2187 predictive time bins across 560 neurons (Extended Data Fig. 7b, d) that were significantly predictive if a Phee call was part of a turn-taking bout (two-sided Mann-Whitney's U-test, $p < 0.05$). Only 763 neurons were considered in these Phee call analyses by the same minimal session selection criteria.

Acknowledgments

This work is supported by grants to AL (Marie Sklodowska-Curie fellowship 101018877 and ERC Starting grant 101116110) and to CTM (NIH R01 DC 012087).

Competing interests

The authors declare no competing interests

Corresponding Author

Correspondence to Arthur Lefevre (arthur.lefevre{at}isc.cnrs.fr).

Supplementary Figure legends

Figure S1. a) Coronal section from the Paxinos atlas showing the location of brush array (top) cannula insertion and Neuropixel (bottom) implantation site. **b)** Coronal MRI planes of monkey S before the first recording session (top) and after the last one (bottom), showing the trajectory of the 64 channels brush array through area 24. **c)** Sagittal MRI plane of monkey G showing (white arrows) the two 32 channels brush arrays implantation sites. **d)** Number of units (total = 2355) recorded per day with the 2 Neuropixel probes implanted in monkey L. Only sessions d4 to d9 were used for data analysis.

Figure S2. a) Picture taken during a recording session. **b)** Table summarizing the number of each call type per caller identity. **c)** Pie charts representing call type proportions for the subject (top) and the partner (bottom). Note that most of the perceived Tsiks and Eks came from a few specific sessions and could not be analyzed. **d)** Examples of spectrograms of recorded vocalizations.

Figure S3. a) Normalized activity of neurons with significant increased response to vocalization production, per call type. **b)** Normalized activity of neurons with significant decreased response to vocalization production, per call type. **c)** Normalized activity of neurons with significant increased response to partner vocalization perception, per call type. **d)** Pie chart representing the neurons (n = 1599) responses: 217 neurons with increased activity for 1 type of vocalization production, 47 neurons with increased activity for several types of vocalization production, 482 neurons with decreased activity for 1 type of vocalization production, 170 neurons with decreased activity for several types of vocalization production, 116 neurons with increased activity for 1 type of partner vocalization perception, 15 neurons with increased activity for several types of partner vocalization perception, 93 neurons with increased activity for 1 type of colony vocalization perception, 51 neurons with increased activity for both vocalization production and partner vocalization perception, 408 neurons with no significant responses.

Figure S4. a) Decoding partner call type accuracy for each session (gray lines) depending on the number of neurons removed, red line is the average. **b)** Percentage accuracy of the neuronal multiclass decoder for perceived call type for real data and data with shuffled labels, after removing neurons with significant activity increase around call onset (Wilcoxon signed rank test, $z = 3.65$, $p < 0.05$). Performance was not significantly different than when significant neurons were included (Mann-Whitney U test, $U = 0.47$, $p > 0.05$). Error bars indicate s.e.m. (standard error to the mean).

Figure S5. a) Decoding accuracy expressed in fold above chance for produced and perceived vocalizations. * indicate significant decoding compared to chance (Wilcoxon sign rank test, $p < 0.05$, Bonferroni corrected). **b)** Correlation between the number of neurons and the decoding quality. **c)** Decoding produced call type performance for each session (gray lines) depending on the number of neurons removed, red line is the average. **d)** Percentage accuracy of the neuronal multiclass decoder for produced call type for real data and data with shuffled labels, after removing neurons with significant activity increase around call onset (Wilcoxon signed rank test, $z = 4.95$, $p < 0.05$). Performance was not significantly different than when significant neurons were included (Mann-Whitney U test, $U = 1{,}61$, $p > 0.05$). Error bars indicate s.e.m. (standard error to the mean).

Figure S6. a) PSTH examples and spike rasters of significant time bins predicting whether a trill call was part of an exchange (within 5 sec of a trill from the other monkey). **b)** PSTH examples and spike rasters of significant time bins predicting whether a phee call was part of an exchange (within 5 sec of a phee from the other monkey). **c-d)** Receiver Operator Characteristic (ROC) traces for our neural decoder trained on predictive time bins and tested on data with shuffled labels representing whether trills answered trills with the subject's partner in panel **c**, the answering of phee calls with the colony in panel **d**. Blue solid lines represent calls initially produced by the subject while orange solid lines represent calls initially perceived by the subject. Random chance is represented by the red dashed line as a guide. The area under the curve (AUC) quantifies the success of the neural decoder as our outcome measure.

Fig 1. *Segregated ACC neuronal populations encode vocalizations perceived from the partner and the colony*

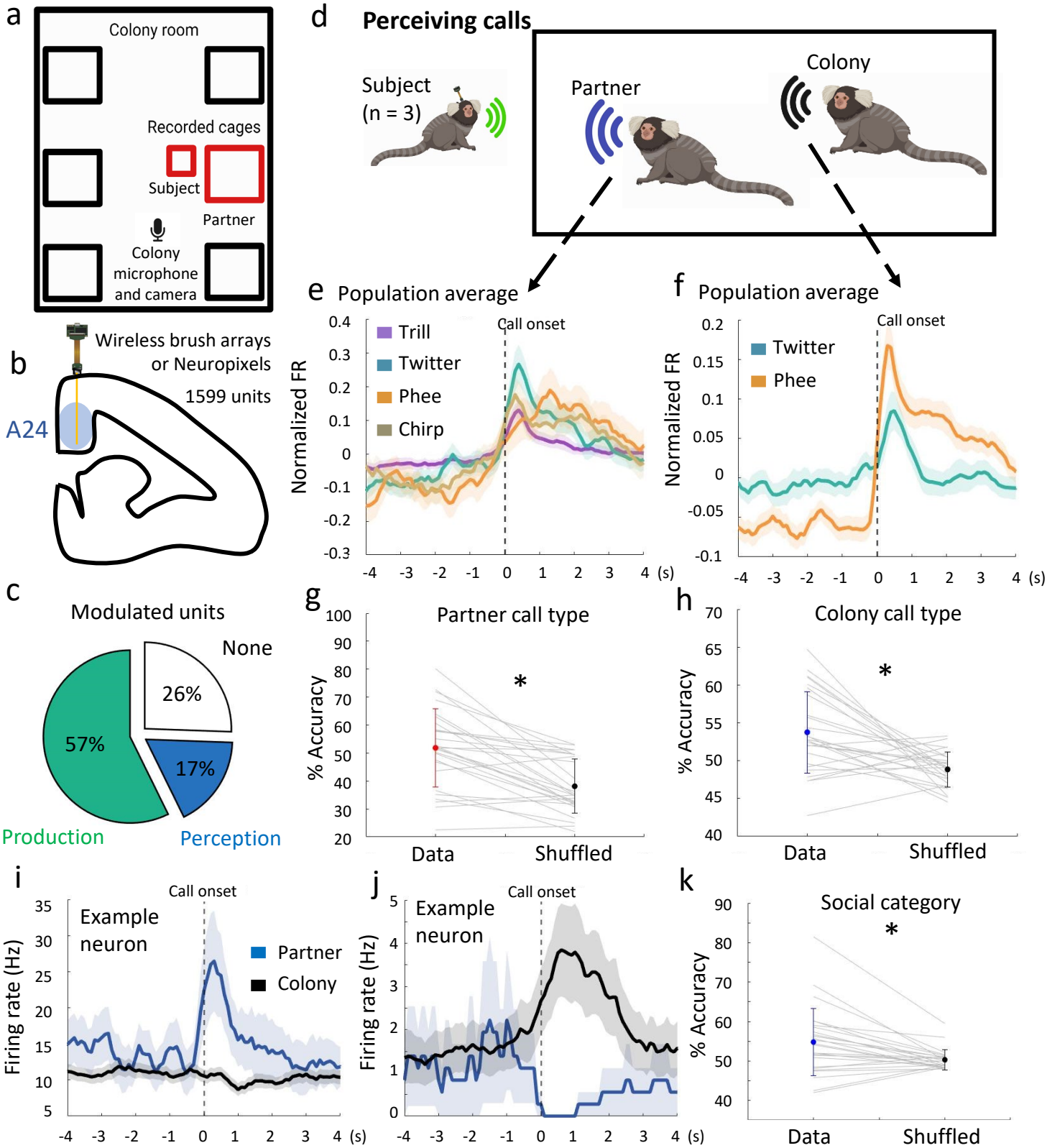


Figure 1. ***Segregated ACC neuronal populations encode vocalizations perceived from the partner and the colony.*** **a)** Schema of the experimental paradigm. *Top* – recordings took place directly in the home cage and transfer box in the colony room. *Center* - both subject and partner monkeys were equipped with a portable microphone, and the implanted subject was also equipped with a neurologger. *Bottom* – Neurons were recorded with brush arrays or Neuropixel probes implanted in ACC area 24. **b)** Coronal plane (AP = +12mm) at which probes were implanted. **c)** Percentage of neurons responding to vocalization production and perception. **d)** Vocalizations produced by the partner or other marmosets in the colony (vs produced ones) were analyzed separately. **e)** Normalized average increase from neurons with significantly upregulated activity during partner call perception. **f)** Normalized average increase from neurons with significantly upregulated activity during colony call perception. **g-h)** Percentage accuracy of the neuronal multiclass decoder for partner **g)** and colony **(h)** perceived call type for real data and data with shuffled labels. **i)** Example unit responding specifically to twitter calls from the partner but not from the colony. **j)** Example unit responding specifically to phee calls from the colony but not from the partner. **k)** Percentage accuracy of the neuronal SVM decoder for identity (partner vs colony) for real data and data with shuffled labels. *: $p < 0.05$, error bars indicate s.e.m, (standard error to the mean).

Fig 2 *ACC units responds invariantly to calls with and without overlapping vocalizations*

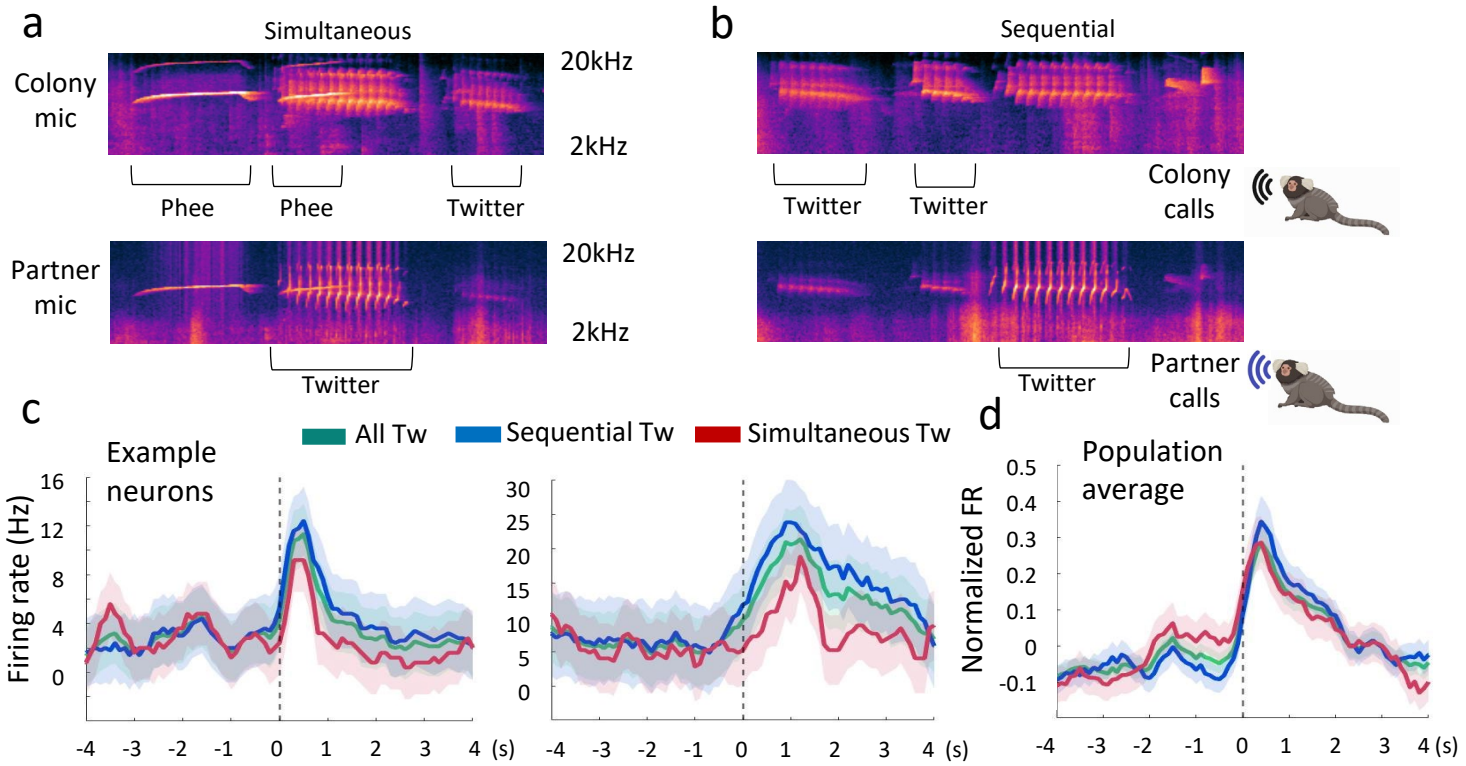


Figure 2. ***ACC units responds invariantly to calls with and without overlapping vocalizations.*** **a)** Example spectrograms from the colony and partner microphones showing the occurrence of a call from the partner (twitter) that is perceived simultaneously with a call from the colony (phee). **b)** Example spectrograms from the colony and partner microphones showing the occurrence of a call from the partner (twitter) that is perceived sequentially to calls from the colony (twitters). **c)** Example neurons responding to twitter calls from the partner that are unaffected by the presence (or absence) of simultaneous calls from the colony. **d)** Normalized average increase from neurons with significantly upregulated activity during partner twitter call perception is unaffected by the presence (or absence) of simultaneous calls from the colony. *: $p < 0.05$, error bars indicate s.e.m.

Fig 3 *Marmosets Monitor Social Scenes to Optimize Vocal Turn Taking in a Cocktail Party*

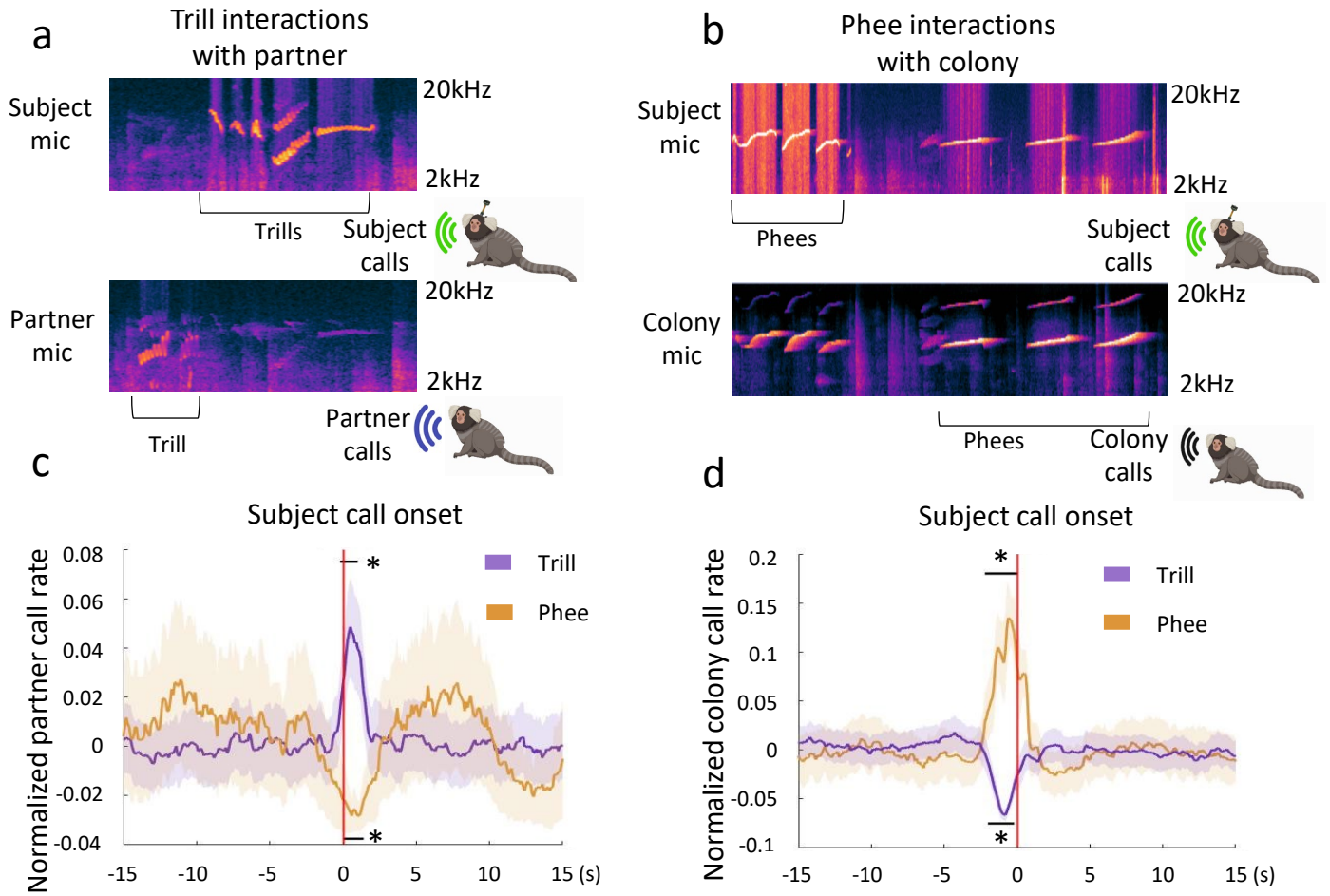


Figure 3. ***Marmosets Monitor Social Scenes to Optimize Vocal Turn Taking in a Cocktail Party.*** **a)** Example spectrograms showing trill to trill conversations between the subject and its partner. **b)** Example spectrograms showing phee to phee conversations between the subject and other monkeys in the colony. **c)** Normalized partner call rate for trills (purple) and phees (orange) aligned on subject's trills and phees onsets. **d)** Colony call rate for phees aligned on subject's trills (purple) and phees (orange).

Fig 4 *ACC neurons encode active turn taking with the partner and the colony*

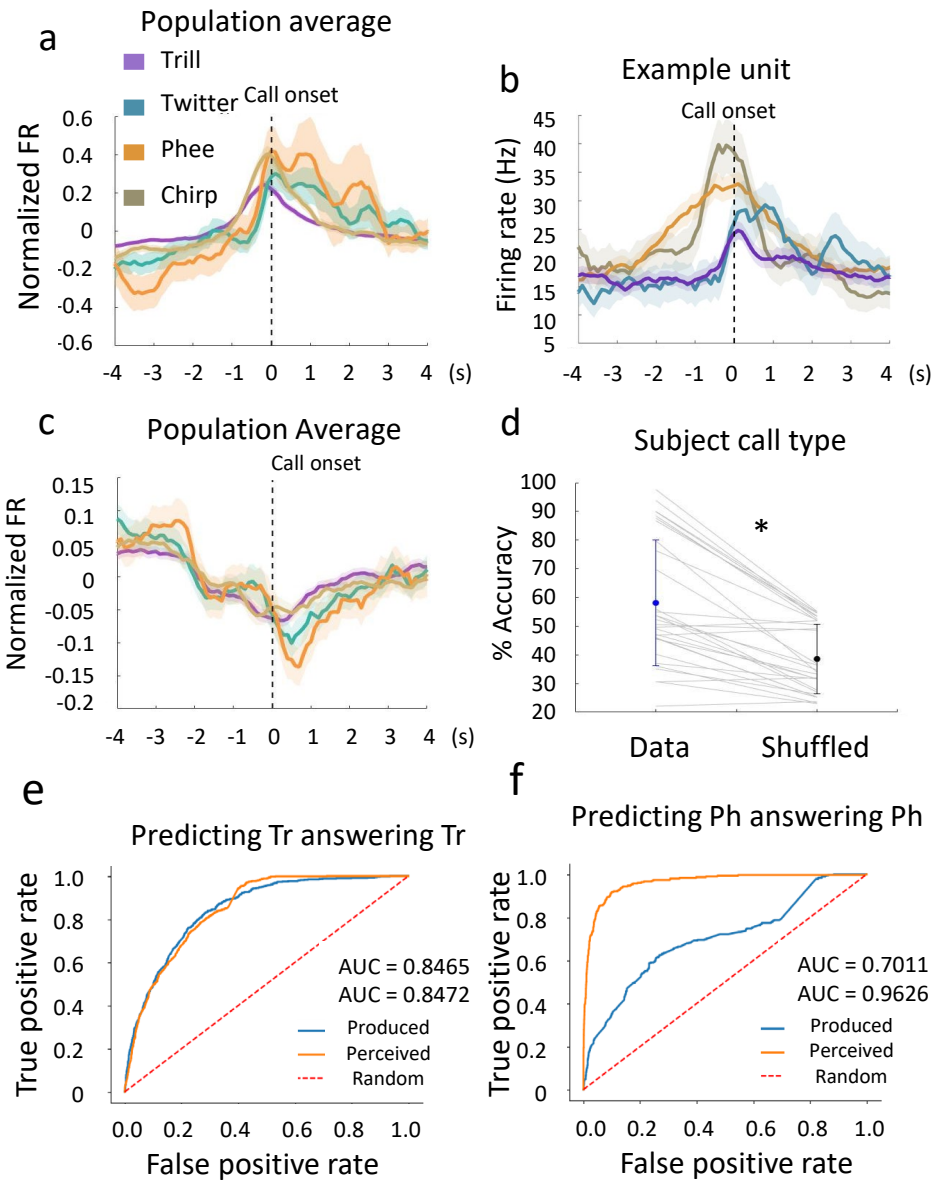


Figure 4. ***ACC neurons encode active turn taking with the partner and the colony.*** **a)** Normalized average increase from neurons with significantly upregulated activity during call production. **b)** Example neuron responding to the production of several call types. **c)** Normalized average decrease from neurons with significantly downregulated activity during call production. **d)** Percentage accuracy of the neuronal multiclass decoder for produced call type for real data and data with shuffled labels. **e-f)** Receiver Operator Characteristic (ROC) traces for our neural decoder trained on predictive time bins representing whether trills answered trills with the subject's partner in panel **e**, the answering of phee calls with the colony in panel **f**. Blue solid lines represent calls initially produced by the subject while orange solid lines represent calls initially perceived by the subject. Random chance is represented by the red dashed line as a guide. The area under the curve (AUC) quantifies the success of the neural decoder as our outcome measure. *: $p < 0.05$, error bars indicate s.e.m.

Figure S1

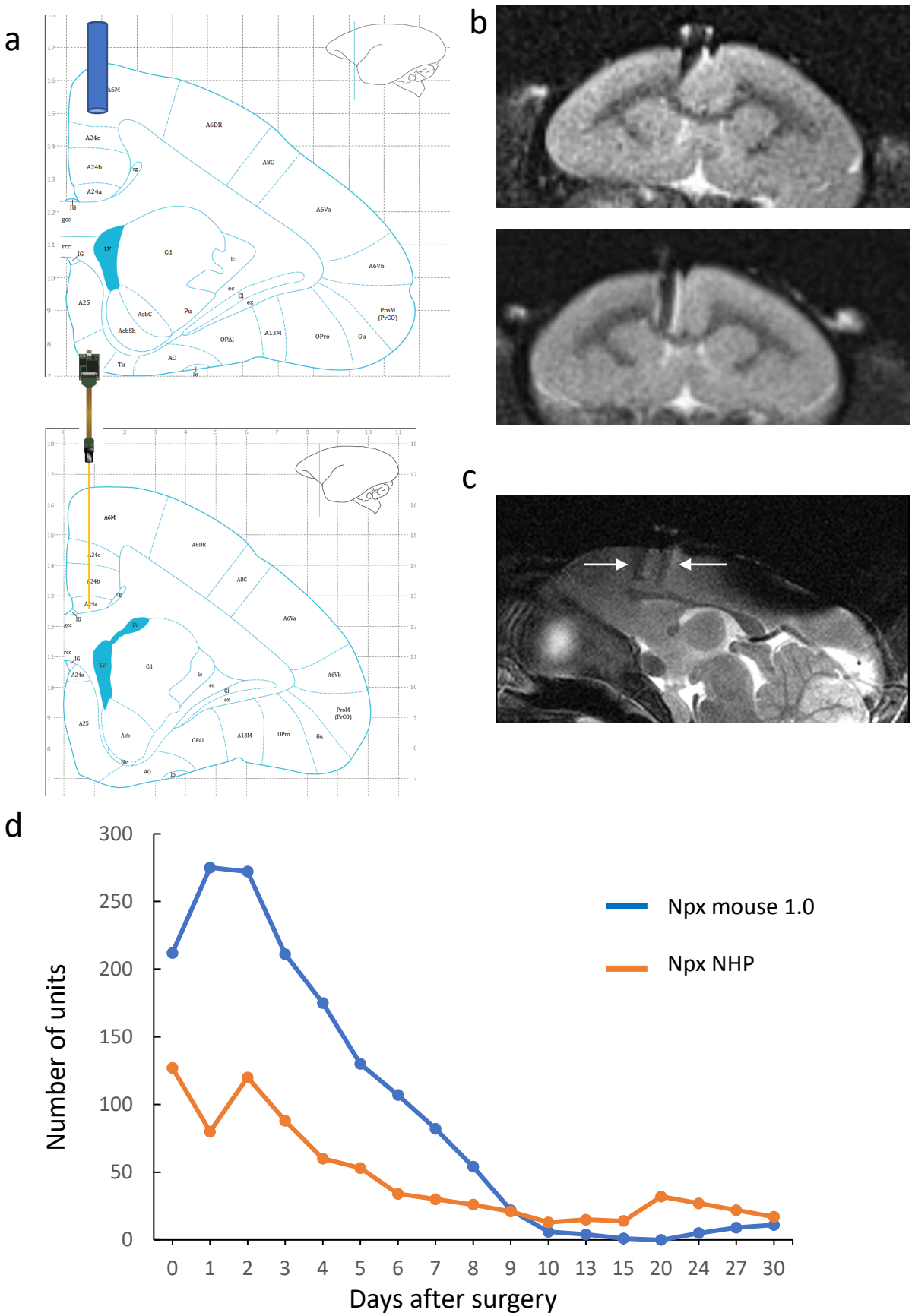
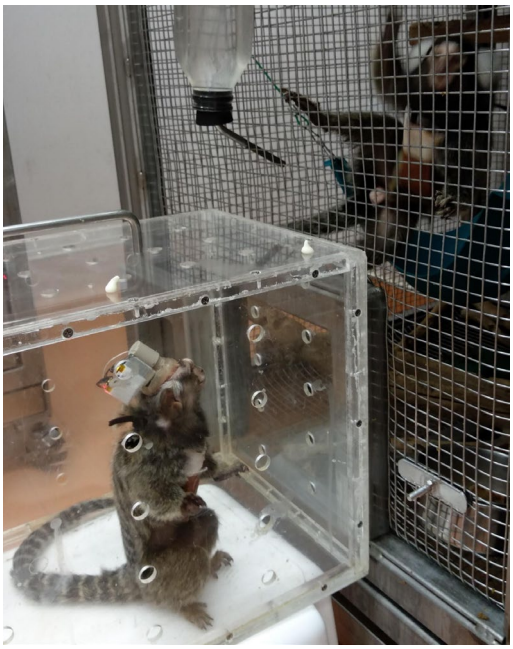


Figure S1. a) Coronal section from the Paxinos atlas showing the location of brush array (top) cannula insertion and Neuropixel (bottom) implantation site. **b)** Coronal MRI planes of monkey S before the first recording session (top) and after the last one (bottom), showing the trajectory of the 64 channels brush array through area 24. **c)** Sagittal MRI plane of monkey G showing (white arrows) the two 32 channels brush arrays implantation sites. **d)** Number of units (total = 2355) recorded per day with the 2 Neuropixel probes implanted in monkey L. Only sessions d4 to d9 were used for data analysis.

Figure S2

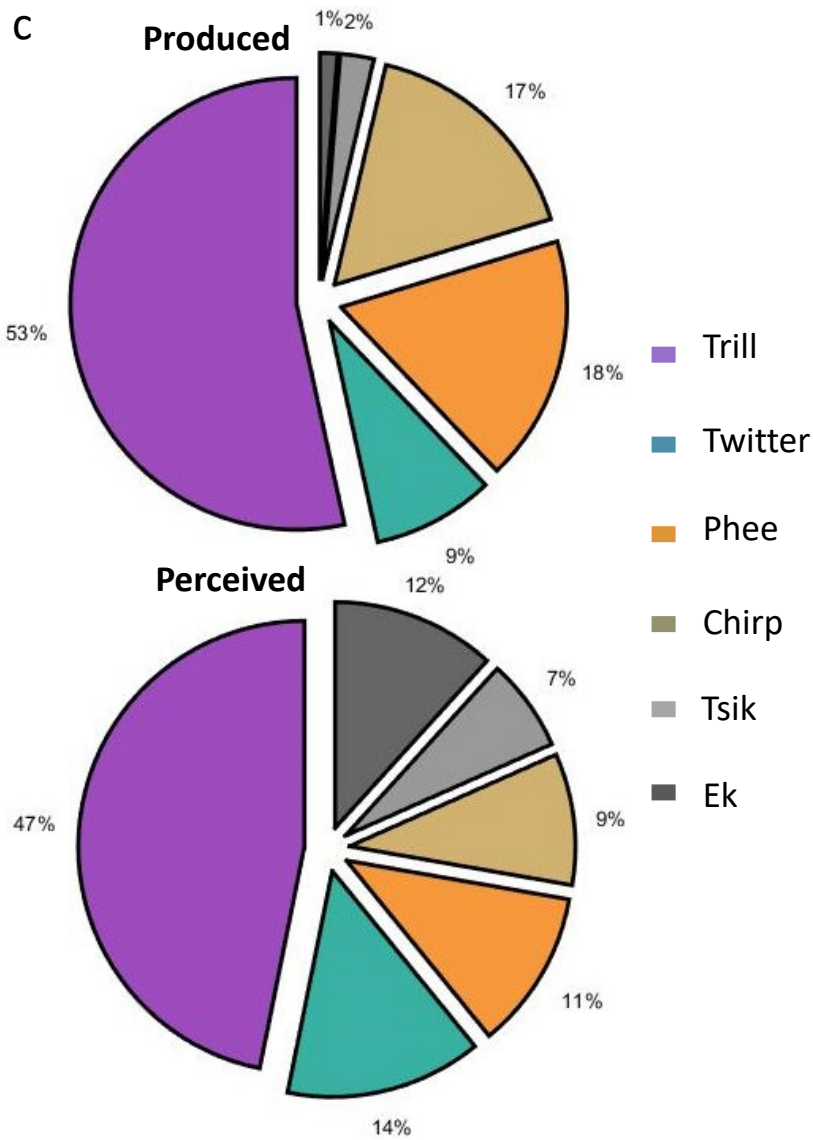
a



b

Voc type	Subject	Partner	Colony
Tr	5386	3201	
Ph	1774	778	12293
Tw	877	959	9350
Chi	1677	643	
Ts	243	452	
Ek	126	802	

c



d

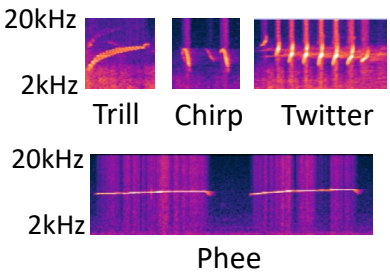


Figure S2. a) Picture taken during a recording session. **b)** Table summarizing the number of each call type per caller identity. **c)** Pie charts representing call type proportions for the subject (top) and the partner (bottom). Note that most of the perceived Tsiks and Eks came from a few specific sessions and could not be analyzed. **d)** Examples of spectrograms of recorded vocalizations.

Figure S3

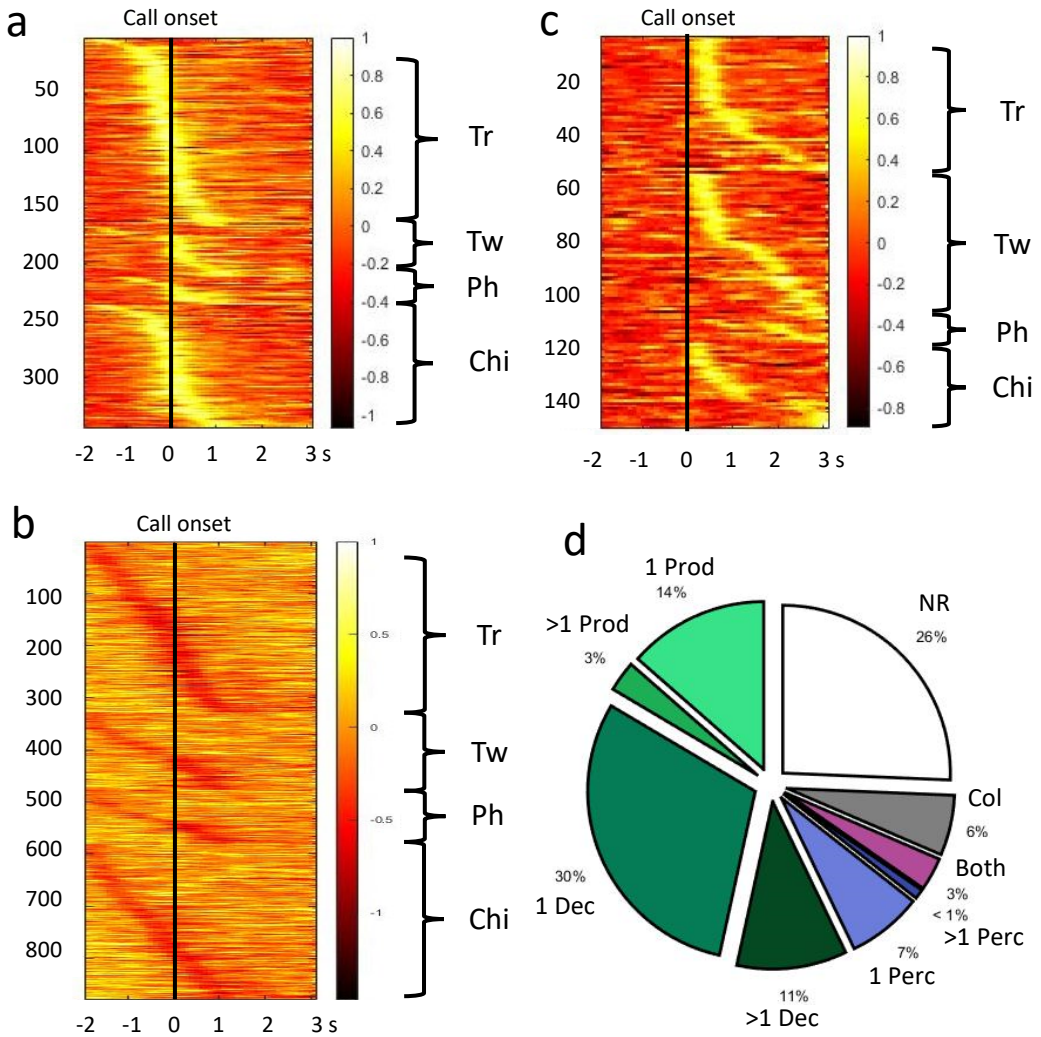


Figure S3. a) Normalized activity of neurons with significant increased response to vocalization production, per call type. **b)** Normalized activity of neurons with significant decreased response to vocalization production, per call type. **c)** Normalized activity of neurons with significant increased response to partner vocalization perception, per call type. **d)** Pie chart representing the neurons (n = 1599) responses: 217 neurons with increased activity for 1 type of vocalization production, 47 neurons with increased activity for several types of vocalization production, 482 neurons with decreased activity for 1 type of vocalization production, 170 neurons with decreased activity for several types of vocalization production, 116 neurons with increased activity for 1 type of partner vocalization perception, 15 neurons with increased activity for several types of partner vocalization perception, 80 neurons with increased activity for 1 type of colony vocalization perception, 51 neurons with increased activity for both vocalization production and partner vocalization perception, 408 neurons with no significant responses.

Figure S4

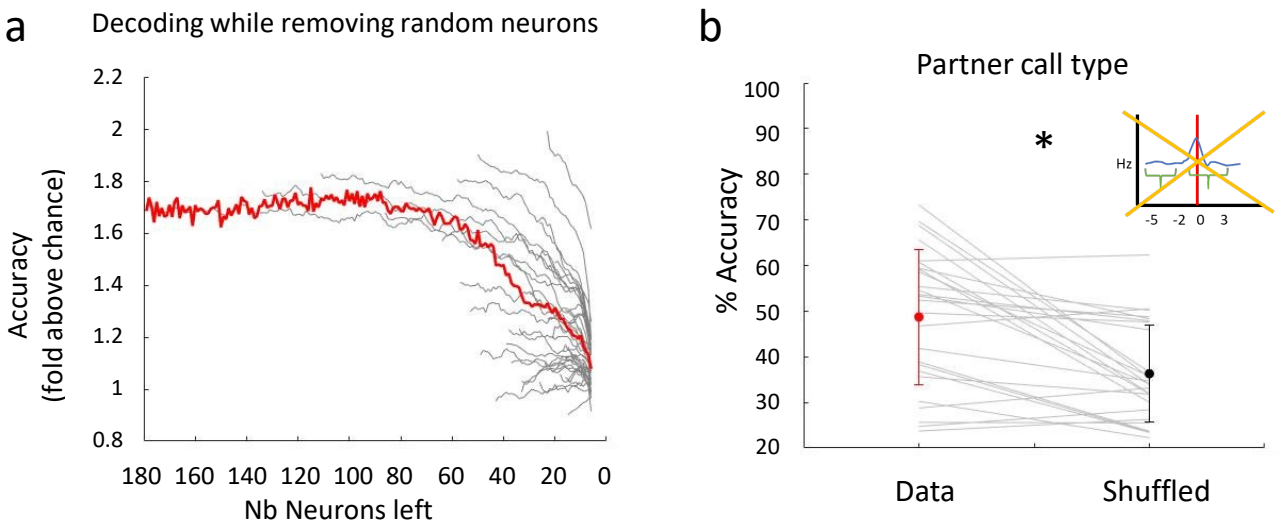


Figure S4. a) Decoding partner call type accuracy for each session (gray lines) depending on the number of neurons removed, red line is the average. **b)** Percentage accuracy of the neuronal multiclass decoder for perceived call type for real data and data with shuffled labels, after removing neurons with significant activity increase around call onset (Wilcoxon signed rank test, $z = 3.65$, $p < 0.05$). Performance was not significantly different than when significant neurons were included (Mann-Whitney U test, $U = 0.47$, $p > 0.05$). Error bars indicate s.e.m. (standard error to the mean).

Figure S5

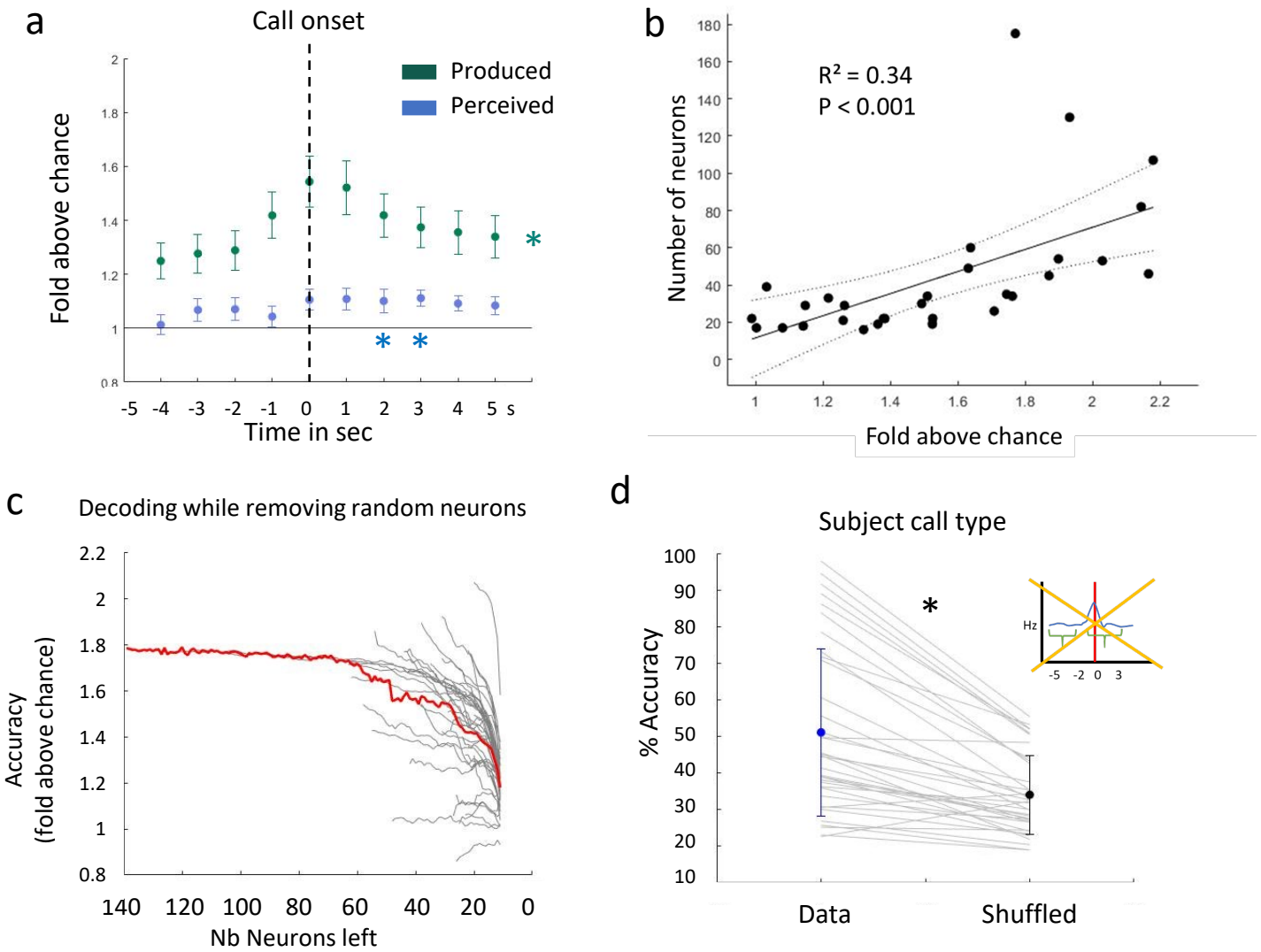


Figure S5. a) Decoding accuracy expressed in fold above chance for produced and perceived vocalizations. * indicate significant decoding compared to chance (Wilcoxon sign rank test, $p < 0.05$, Bonferroni corrected). **b)** Correlation between the number of neurons and the decoding quality. **c)** Decoding produced call type performance for each session (gray lines) depending on the number of neurons removed, red line is the average. **d)** Percentage accuracy of the neuronal multiclass decoder for produced call type for real data and data with shuffled labels, after removing neurons with significant activity increase around call onset (Wilcoxon signed rank test, $z = 4.95$, $p < 0.05$). Performance was not significantly different than when significant neurons were included (Mann-Whitney U test, $U = 1{,}61$, $p > 0.05$). Error bars indicate s.e.m. (standard error to the mean).

Figure S6

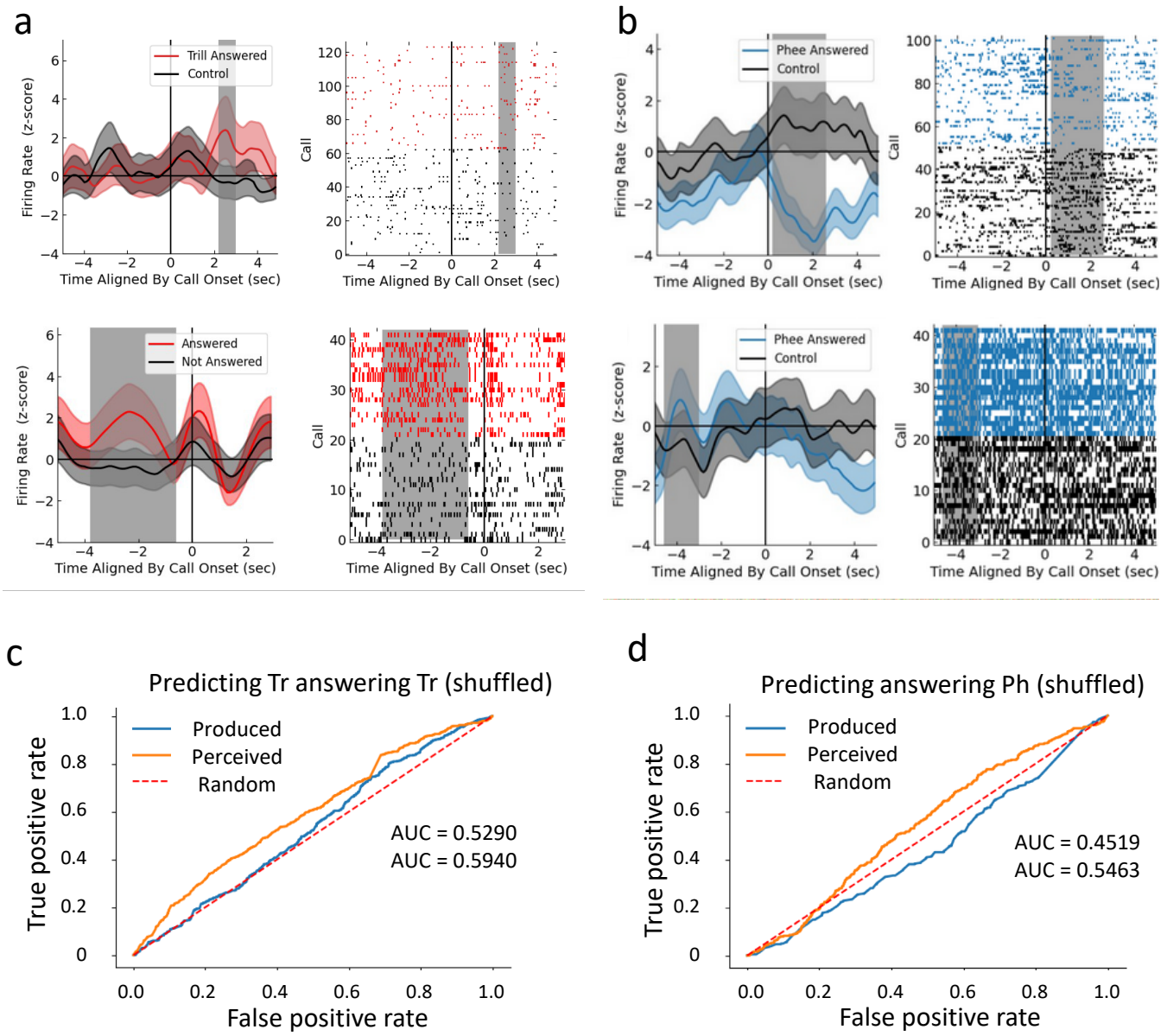


Figure S6. a) PSTH examples and spike rasters of significant time bins predicting whether a trill call was part of an exchange (within 5 sec of a trill from the other monkey). **b)** PSTH examples and spike rasters of significant time bins predicting whether a phee call was part of an exchange (within 5 sec of a phee from the other monkey). **c-d)** Receiver Operator Characteristic (ROC) traces for our neural decoder trained on predictive time bins and tested on data with shuffled labels representing whether trills answered trills with the subject's partner in panel **c**, the answering of phee calls with the colony in panel **d**. Blue solid lines represent calls initially produced by the subject while orange solid lines represent calls initially perceived by the subject. Random chance is represented by the red dashed line as a guide. The area under the curve (AUC) quantifies the success of the neural decoder as our outcome measure.