Title

Cortical Communication in the Context of Learning

Author

Veuthey, Tess

Publication Date

2019

Cortical Communication in the Context of Learning

by
Tess Veuthey

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Neuroscience

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Loren Frank — Loren Frank
EF21CD5A6EE643A...
Chair

DocuSigned by:
Michael Brainard — Michael Brainard
[illegible]412...

DocuSigned by:
Massimo Scanziani — Massimo Scanziani
[illegible]404...

DocuSigned by:
Karunesh Ganguly — Karunesh Ganguly
537133AC2605407...

Committee Members

To my parents, siblings, family, and communities,
whose support made this possible.

Acknowledgements

First and foremost, I would like to thank my thesis advisor, Karunesh Ganguly, for his guidance and support. Three of his qualities that were particularly important to my PhD experience are his enthusiasm, optimism, and insight. Karunesh finds a way to be excited about the myriad of projects in his lab throughout their ups and downs, which was critically important during my moments of cynicism or discouragement. He additionally was crucial in teaching me to frame my work in the larger context of systems neuroscience. I am certain that graduate school would have been less enjoyable without him as a guide.

My other thesis committee members, Loren Frank (chair), Michael Brainard, and Massimo Scanziani, also offered invaluable advice. Loren Frank consistently reminded me to be skeptical of needlessly complex analyses. Michael Brainard always had wisdom about designing experiments with the purpose of distinguishing multiple hypotheses. Massimo both suggested incisive analyses and reminded me to frame the work in terms of fundamental questions about the nervous system. All of them have offered anecdotes and advice about life outside of lab, which for me is key for mentoring students holistically. I also want to thank Flip Sabes for the advice he offered before leaving UCSF. Alexandra Nelson and Vikaas Sohal were additionally helpful in the initial shaping of my project as members of my qualifying exam committee.

In lab, I give special thanks to my outstanding collaborator Kate Derosier, who joined our lab when the M1-M2 Reaching project was already underway, and who helped shape it in a more interesting computational direction. Kate’s ability to learn and teach relevant mathematical approaches, her scientific integrity, and her overall patience and perceptiveness have been invaluable. In general, the Ganguly Lab members are fantastic colleagues and scientific conversation partners. In particular, Dhakshin Ramanathan and Tanuj Gulati trained me to do electrophysiological implants; Daniel Silversmith had excellent feedback on project framing; Preeya Khanna was essential for helping me understand Factor Analysis and its scientific context; and Jaekyung Kim and Sravani Kondapavulur were great rig-sharing companions. Seok Joon Won and Chelsey Campillo Rodriguez provided help with histology and imaging. Finally, Ling Guo, Stefan Lemke and I entered lab as a cohort, and they have been my most constant companions in the Ganguly Lab; the experience would have been more daunting and less fun without their company. Our shared experiences cajoling rats, battling recording equipment, and experimenting with new surgical techniques have cemented a lasting bond.

I would also like to thank the crew at 1700 Owens as a whole, including the LARC staff and the front desk staff, especially Anthony and Frank, whose welcome each morning set the tone for the day. Finally, thanks to all the rats who participated in this work.

At UCSF, I would like to thank the Neuroscience Graduate Program, particularly Anatol Kreizer, Roger Nicoll, Pat Veitch, and Lucita Nacionales. The 2014 entering class have been wonderful companions, especially Michael Ryan and Nerissa Hoglen, who provided constant support and scientific feedback. Other science friends at UCSF and elsewhere with whom I had formative conversations about science and graduate school include Allison Girasole, Mari Sosa, Lucas Tian, Anna Gillespie, Riccardo

Beltramo, Hamish Mehaffey, Jai Yu, Demetris Roumis, Kurtis Swartz, Will Liberti, Daniel Estandian, Sama Ahmed, Tyler Bonnen, Emily Mackevicius, and Thomas O'Connell.

The UCSF MSTP has also been a constant source of perspective. Thanks to the leadership and administrators, Mark Anderson, Kevin Shannon, Geri Ehle, Jana Toutolmin, and Catherine Norton. The mentorship of Daniel Lowenstein and Alexandra Nelson have been priceless from the beginning of my time in the program. Other faculty members who provided career and clinical mentoring include Aimee Kao, Cathy Lomen-Hoerth, and David Kastner. 'Elder' students who were rich sources of guidance include Alex Ramos, Anthony Lee, David Staudt, Jason Chung, Anna Hauswirth, and Kenneth Kay. Thanks also to Hannah Joo and Jen Guidera. Big thanks to Lay Kodama, Parinaz Fozouni, Chimno Nnadi, and Joseph Hiatt for their work on improving the experiences of women and minorities in the UCSF MSTP. Finally, thanks to the 2012 entering class for their constant humor and companionship, especially Max Liu, my classmate in both the MSTP and the Neuroscience Graduate Program, who has spent many hours sitting next to me while I made thesis committee slides or otherwise figured out the next steps in this long process. Finally, thanks to the UCSF School of Medicine and the larger UCSF community for creating a wonderful environment in which to train. This includes Jess Kim, Ally Girasole, Ally Wong, and Stefan Lemke for helping me train in open-water swimming to participate in the Alcatraz crossing in October 2018. Thanks to so many learning companions in the School of Medicine, including Annie Hargrave, Colette DeJong, Marie Sears, Sharon Stortz, Libby McNiven, Adam Tapley, Sarra Borhanian, Leslie Suen, Stacey Frumm, and others. Thanks to Samuel Thompson for being a great writing companion in addition to an honest friend and believer in ‘direct-yet-loving’ criticism.

Thanks to my previous neuroscience mentors, particularly Nancy Kanwisher, Mark Baxter, and Kathy Murphy. Finally, the communities at the Redwood Center for Theoretical Neuroscience and the Center from Brains, Minds, and Machines (at Woods Hole), who created fun and enlightening summer course experiences.

Thanks to many non-science friends outside of UCSF. To Jen Kyhos for being patient with my workaholism these past 20 years. To Elysa Wan, Jess Schimer, and Maria Luckyanova, for our regular video-chats which always provide perspective on the world at large. Thanks for my ‘beau-frère’ Prashant Kotwani, for being a great roommate. Thanks to the members of Book Club and Docu-Club, for providing regular communities to delve into new worlds and topics together. Thanks to my exes for providing support during different phases of my training. Thanks to Rachel Faller, Radu Raduta, Felix Fang, Chris Magoufis, Constantine Firinidis, Constantinos Melachrinos, Alan Leonard, Lluis Peñalver, Marine Ray, Khoa Vu, Theo Rodman, and so many others. Finally, thanks to Prasad Shirvalkar, for being an understanding and supportive partner these past few years. His unwavering faith in my scientific abilities and excitement about neuroscience have kept me afloat after long nights in lab.

Thanks to my paternal and maternal families, based in Geneva and the Bay Area, who have always expressed curiosity and support for my work. Thanks to my siblings, Claire and Justin, for their constant encouragement and for being great sources of inspiration throughout my life. Finally, thanks to my parents, Cora Shaw and

Bernard Veuthey. There are no words to express my love and gratitude for their wisdom and generosity over the years.

This research was conducted with government funding under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. Additional funding came from the UCSF Medical Scientist Training Program, the UCSF Discovery Fellows Program, and the Markowski-Leach Fellowship.

Contributions

Chapter 3 is reprinted largely as it appears in: Shirvalkar, P., Veuthey, T. L., Dawes, H. E. & Chang, E. F. Closed-Loop Deep Brain Stimulation for Refractory Chronic Pain. Front. Comput. Neurosci. 12, (2018).

Appendix A is reprinted largely as it appears in: Perkovich, B. & Veuthey, T. L. Do Ask, Do Tell: UCSF SOM asks applicants about sexual orientation and gender identity. (2014). Available at: https://synapse.ucsf.edu/articles/2014/07/31/do-ask-do-tell-ucsf-som-asks-applicants-about-sexual-orientation-and-gender.

Appendix B is reprinted largely as it appears in: Gates, M. A. et al. Tiptoeing around it: Inference from absence in potentially offensive speech. in CogSci (2018).

Appendix C is reprinted largely as it appears in: Veuthey, T. L. & Thompson, S. Why you need an agenda for meetings with your principal investigator. Nature 561, 277 (2018).

Cortical Communication in the Context of Learning

Tess L. Veuthey

Abstract

The infinite range of human behaviors is made possible by the anatomic and functional complexity of our brains. Our brains are arranged as networks of interacting neural populations which perform computations both within and across areas. Past research has focused on the specific roles of different brain regions, parceling out computational steps in sensory, motor, cognitive, and affective processes. However, our understanding of how brain regions interact is extremely preliminary, and is bottlenecked by limitations in experimental approaches, recording technologies, interventional methods, and computational analyses. These limitations impact not only our comprehension of the nervous system, but also our ability to design, optimize, and implement new therapies for patients with neurological diseases and disorders.

This thesis first investigates cross-area communication in the motor system in the context of natural movement learning and closed-loop brain-machine interface (BMI) learning. It then proposes a framework for understanding and manipulating cross-area communication in the context of chronic pain, a disorder driven by pathological activation and coupling of sensory, cognitive, and affective regions. We find that, during natural motor learning, cross-area activity dynamics can (1) be distinguished from local dynamics, (2) develop representations of learned movements which predict single-trial behavior, (3) become coordinated with local dynamics over the course of learning, and (4) causally influence downstream local activity to drive learned behaviors. Preliminary results from BMI learning show that tasks requiring only local neural modulation also engage neural populations in partner brain regions, suggesting circuit-wide participation in new task learning. This knowledge of the brain’s ability to learn new cross-area activity patterns in the context of natural behaviors and external, device-based feedback informs our framework for designing closed-loop neuromodulatory therapies for refractory chronic pain given the devices currently available for treating nervous system disorders.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Introduction

The brain is a classic example of a complex system: it is made of relatively simple individual agents, neurons, whose diverse and widespread interactions lead to an infinite range of behaviors[1,2]. The neural connections necessary for supporting life are set up during development, but neural interactions must continue evolving in adult animals in order to support learned behaviors[3–5]. While the evolution of neural interactions within a brain region have been extensively studied, relatively little is known about how neural interactions between brain regions change with learning. This dearth of knowledge is in part due to the intractable nature of the brain as a system with billions of neurons and trillions of neural connections. To reduce the scope of the problem, research has focused on the role of individual brain regions in during specific behaviors.

Early neurological reports of patients with localized brain lesions and specific behavioral deficits led to the realization that neurons are functionally organized into distinct brain regions [6–9]. Brain regions, in turn, are connected into networks supporting sensory, motor, cognitive, and affective functions. In order to explicitly study neural functions rather than relying on patients' spontaneously occurring brain damage and ensuing deficits, neuroscientists use animal models to probe neural activity underlying carefully designed behaviors. Often animals are first extensively trained on tasks designed to probe specific behavioral parameters [10]; then neural activity from a single brain region is either (a) recorded and analyzed to discover correlations with behavioral parameters [11,12] or (b) disrupted to discover the necessity of its function in that behavior [13,1412]. This approach has yielded rich knowledge on how activity restricted to single brain regions underlies expert performance in a task. However, it has been difficult to expand this approach to understand how brain regions interact in the context of learning, especially at the level of neural populations. This is due to (1) the technological limitations for obtaining simultaneous data from neural populations in multiple brain regions in a behaving animal, (2) a sparsity of established methodologies for relating cross-area neural activity, and (3) difficulty in interpreting neural activity related to variable behavior. However, recently, multi-site neural population recording technologies [15,16], computational dimensionality reduction methods [17–19], and approaches to single-trial neural data interpretations [20–25] have emerged as candidates for making analysis of cross-area communication during learning a tractable problem.

Chapter 1 uses (1) simultaneous neural population recordings in premotor (M2) and primary (M1) motor cortex, (2) a combination of dimensionality reduction methods designed to extract neural signals either local[4,26,27] to a brain region or shared[28,29] across two brain regions, and (3) single-trial neural analyses to understand how M1-M2 cross-area communication evolves to support a learned motor skill. Key to this study was the novel use of Canonical Correlation Analysis [28] (CCA) in the investigation of simultaneously recorded population activity from two brain regions. As outlined above, CCA is a dimensionality reduction method designed to detect and extract maximally correlated information across two sets of signals. We compared these cross-area neural signals to those extracted using Factor Analysis (FA)[26,27], a dimensionality reduction method designed to detect variance that is shared within a single population of signals. This approach allowed us to track the relationship between neural signals defined locally in M1 or M2 (i.e. local dynamics), and signals defined by activity that M1 and M2 have in common (i.e. cross-area dynamics). The additional use of single-trial analyses allowed us to examine neural activity during variable learning behavior.

By combining these strategies, we found that emergence of coordination between local and cross-area population dynamics drives learned motor behaviors. We tested the necessity of coordinated M1-M2 activity by inactivating M2 in well-trained animals. M2 inactivation resulted in both behavioral deficits and disruption of M1's ability to encode learned movements. Importantly, neither the behavior nor M1's encoding of movement were completely abolished, demonstrating local resilience in M1 to a distant disruption within the functional motor network (here M2). These findings and others in this study indicate that evolving interactions both within and between nodes of the motor network can be probed to understand neural correlates of natural motor learning.

Chapter 2 addresses a major limitation inherent in the study of natural movements. Namely, that many motor area have extensive bi-directional connections to each other as well as parallel connections to downstream regions. Consequently, it is often impossible to claim that one any region uniquely controls a particular parameter of movement, whether it be abstract (e.g. reaction time) or kinematic (e.g. hand shaping during grasping). For example, in rodents, forelimb regions in M1 and M2 have dense cross-area connections and both send projections to the same segments of spinal cord [30]. Descriptions of M1 and M2's specific roles are dependent on the task in which they are probed, leading to results suggesting that M2 contains more signals related to movement context than M1 [31]; that M2 is more related to distal grasping movements than M1[32]; that M2 and M1 have opposite influence on the near versus far reach targeting [33]; and M1 and M2 have parallel, but nearly identical functions [34,35]. These conclusions are not mutually exclusive, and they highlight the limitations to understanding directionality of cross-area communication during natural movements.

To address this limitation, we can use brain-machine interfaces [36] (BMIs) to specifically constrain and design the relationship between neural activity and effectors. BMIs allow us to directly map activity from selected neurons into signals that control artificial effector movements. By design, the activity of all other neurons is not required for the artificial effector movements. Consequently, any relationship of those neurons to the task is due either to inherent functional neural connectivity, and/or the animal's inability to distinguish which neural signals are necessary for the task. To examine the inherent functionally connectivity of M2 and M1, we simultaneously recorded in M2 and M1 during a M1-driven BMI task. We found that M2 neurons are driven by M1-BMI learning, suggesting that M1-M2 cross-area connections are engaged during M1-BMI learning.

Chapter 3 discusses how inherent cross-cortical communication can become pathological, and how neuromodulatory interventions can be used to therapeutically decouple cross-cortical communication. Specifically, this chapters frames chronic pain as pathological coupling between areas involved in the sensory, cognitive, and affective components of pain [37–39], leading to resonant circuit activity and recurrent entry into brain states associated with pain (i.e. pain state). We outline how four types of deep brain stimulation (DBS) might be used to treat chronic pain by disrupting communication between regions. Inherent aspects of the four types of DBS, (1) single-site open-loop DBS, (2) patient-triggered on/off DBS, (3) sensor-triggered on/off DBS, and (4) multi-site closed-loop DBS lead to very different goals for DBS-based neuromodulation. In short, since single-site open-loop DBS cannot monitor the patient's brain state, the goal of therapy must be to permanently keep the underlying functional network out of the pain state, potentially increasing the risk of side effects resulting from decoupling within the functional network. In contrast, patient-triggered DBS relies on a patient-detectable level of pain, and consequently is designed to abort pain rather than avoid it. Similarly, sensor-triggered DBS would be designed to avoid crossing a pre-determined threshold in the brain state representation, leading to restrictions in accessibility of brain states. Finally, multi-site closed-loop DBS has the potential to titrate area-specific and cross-area neurostimulation to prevent entry into the global pain state without constraining the overall variability of neural activity within each brain region.

Overall, this thesis (1) provides evidence for the evolution of cross-area interactions within a functional network during natural movement learning; (2) highlights that cross-area communication is intrinsically engaged in task learning, even when it is not apparently required for task execution; and (3) proposes a strategy for manipulating functional networks when cross-area interactions lead to pathological communication.

References

1. Sporns, O., Chialvo, D., Kaiser, M. & Hilgetag, C. Organization, development and function of complex brain networks. *Trends Cogn. Sci.* **8**, 418–425 (2004).

2. Avena-Koenigsberger, A., Misic, B. & Sporns, O. Communication dynamics in complex brain networks. *Nat. Rev. Neurosci.* **19**, 17–33 (2018).

3. Cao, V. Y. *et al.* Motor Learning Consolidates Arc-Expressing Neuronal Ensembles in Secondary Motor Cortex. *Neuron* (2015). doi:10.1016/j.neuron.2015.05.022

4. Athalye, V. R., Ganguly, K., Costa, R. M. & Carmena, J. M. Emergence of Coordinated Neural Dynamics Underlies Neuroprosthetic Learning and Skillful Control. *Neuron* **93**, 955-970.e5 (2017).

5. Nudo, R. J., Milliken, G. W., Jenkins, W. M. & Merzenich, M. M. Use-dependent alterations of movement representations in primary motor cortex of adult squirrel monkeys. *J. Neurosci.* **16**, 785–807 (1996).

6. Harlow, J. Passage of an iron rod through the head, Boston M. & S. *J* **39**, 389 (1848).

7. Harlow, I. *Recovery from the Passage of an Iron Bar through the Head. Publications ofthe Massachusetts Medical Society 2, 327–347*. (M. Mcmil—lan (2002), An odd kind offame. Stories ofPhineas Gage. London …, 1868).

8. Fye, W. B. Julien Jean César Legallois. *Clin. Cardiol.* **18**, 599–600 (1995).

9. Ferrier, D. The Goulstonian Lectures on the Localisation of Cerebral Disease. *Br. Med. J.* **1**, 443–447 (1878).

10. Kawai, R. *et al.* Motor Cortex Is Required for Learning but Not for Executing a Motor Skill. *Neuron* **86**, 800–812 (2015).

11. Kay, K. *et al.* A hippocampal network for spatial coding during immobility and sleep. *Nature* (2016). doi:10.1038/nature17144

12. Beltramo, R. & Scanziani, M. A collicular visual cortex: Neocortical space for an ancient midbrain visual structure. *Science* **363**, 64–69 (2019).

13. Tian, L. Y. & Brainard, M. S. Discrete Circuits Support Generalized versus Context-Specific Vocal Learning in the Songbird. *Neuron* **96**, 1168-1177.e5 (2017).

14. Gulati, T., Guo, L., Ramanathan, D. S., Bodepudi, A. & Ganguly, K. Neural reactivations during sleep determine network credit assignment. *Nat. Neurosci.* **advance online publication**, (2017).

15. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).

16. Chung, J. E. *et al.* High-Density, Long-Lasting, and Multi-region Electrophysiological Recordings Using Polymer Electrode Arrays. *Neuron* **101**, 21-31.e5 (2019).

17. Pang, R., Lansdell, B. J. & Fairhall, A. L. Dimensionality reduction in neuroscience. *Curr. Biol.* **26**, R656–R660 (2016).

18. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).

19. Laubach, M., Wessberg, J. & Nicolelis, M. A. L. Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. *Nature* **405**, 567–571 (2000).

20. Churchland, M. M., Yu, B. M., Sahani, M. & Shenoy, K. V. Techniques for extracting single-trial activity patterns from large-scale neural recordings. *Curr. Opin. Neurobiol.* **17**, 609–618 (2007).

21. Kao, J. C. *et al.* Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nat. Commun.* **6**, 7759 (2015).

22. Wei, Z., Inagaki, H., Li, N., Svoboda, K. & Druckmann, S. An orderly single-trial organization of population dynamics in premotor cortex predicts behavioral variability. *Nat. Commun.* **10**, 216 (2019).

23. Yu, B. M. *et al.* Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *J. Neurophysiol.* **102**, 614–635 (2009).

24. Mackevicius, E. L. *et al.* Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *eLife* **8**, e38471 (2019).

25. Kiani, R., Cueva, C. J., Reppas, J. B. & Newsome, W. T. Dynamics of Neural Population Responses in Prefrontal Cortex Indicate Changes of Mind on Single Trials. *Curr. Biol.* **24**, 1542–1547 (2014).

26. Ghahramani, Z. & Hinton, Geoffrey. The EM Algorithm for Mixtures of Factor Analyzers. 8 (1997).

27. Toutenburg, H. Everitt, B. S.: Introduction to Latent Variable Models. Chapman and Hall, London 1984. 107 pp., £ 9.50. *Biom. J.* **27**, 706–706 (1985).

28. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28**, 321–377 (1936).

29. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical Areas Interact through a Communication Subspace. *Neuron* (2019). doi:10.1016/j.neuron.2019.01.026

30. Rouiller, E. M., Moret, V. & Liang, F. Comparison of the Connectional Properties of the Two Forelimb Areas of the Rat Sensorimotor Cortex: Support for the Presence of a Premotor or Supplementary Motor Cortical Area. *Somatosens. Mot. Res.* **10**, 269–289 (1993).

31. Saiki, A. *et al.* Different Modulation of Common Motor Information in Rat Primary and Secondary Motor Cortices. *PLoS ONE* **9**, e98662 (2014).

32. Brown, A. R. & Teskey, G. C. Motor Cortex Is Functionally Organized as a Set of Spatially Distinct Representations for Complex Movements. *J. Neurosci.* **34**, 13574–13585 (2014).

33. Wahl, A. S. *et al.* Optogenetically stimulating intact rat corticospinal tract post-stroke restores motor control through regionalized functional circuit formation. *Nat. Commun.* **8**, 1187 (2017).

34. Morandell, K. & Huber, D. The role of forelimb motor cortex areas in goal directed action in mice. *Sci. Rep.* **7**, 15759 (2017).

35. Hyland, B. Neural activity related to reaching and grasping in rostral and caudal regions of rat motor cortex. *Behav. Brain Res.* **94**, 255–269 (1998).

36. Fetz, E. E. Operant conditioning of cortical unit activity. *Science* **163**, 955–958 (1969).

37. Melzack, R. From the gate to the neuromatrix. *PAIN* **82**, S121 (1999).

38. Melzack, R. & Casey, K. Sensory, Motivational and Central Control Determinants of Pain: A New Conceptual Model. in *The Skin Senses* (1968).

39. Melzack, R. & Wall, P. D. Pain Mechanisms: A New Theory. *Science* **150**, 971–979 (1965).

Chapter 1: Emergent coordination of local and cross-area population dynamics drives learned motor behaviors

Authors: T. L. Veuthey[1,2,3,4]†, K. Derosier[1,3,4]†, K. Ganguly[3,4].

Affiliations:

[1]Neuroscience Graduate Program, University of California San Francisco, San Francisco CA.

[2]Medical Scientist Training Program, University of California San Francisco, San Francisco CA.

[3]Neurology and Rehabilitation Service, San Francisco Veterans Affairs Medical Center, San Francisco CA, USA.

[4]Department of Neurology, University of California San Francisco, San Francisco CA, USA.

† These authors contributed equally to this work.

Article Type: Original Research

Keywords: Motor Learning, Motor Cortex, Premotor Cortex, Animal Behavior, Electrophysiology

Abstract

Mammalian cortex is a complex system with both local and cross-area connections. The combination of these two motifs suggests a vital role for interactions between local and cross-area neural population dynamics. However, prior work has not distinguished how local versus cross-area activity dynamics might differentially drive learning and skilled execution. Here we hypothesize that interactions between local population dynamics with those that coordinate dynamics across areas are necessary for skilled motor behaviors. Using multisite recordings of motor (M1) and premotor (M2) cortex along with computational modeling, we analyzed how local and cross-area activity patterns interact during reach learning in rats. Strikingly, the emergence of reach-related modulation in cross-area activity appeared to drive skill acquisition. Additionally, the single-trial modulation in cross-area activity was predictive of both reaction time and reach duration. Furthermore, coordination of cross-area dynamics with local dynamics increased significantly with skill learning. Consistent with a functional role for cross-area dynamics, M2 inhibition disrupted both M1 dynamics and reach behavior. Together, these results indicate that coordination of task signals between local and cross-area population dynamics is necessary for skilled motor behaviors.

Introduction

The connectivity pattern of mammalian cortex, characterized by both local and cross-area connections[1], suggests an important role for interactions between population dynamics compartmentalized locally with those coordinated between regions. But it is unknown whether dynamics coordinated across multiple cortical areas contribute to learning and skilled execution. For example, in the motor system, it has been shown that both premotor cortex (M2) [2–6] and motor cortex (M1) [7–11] demonstrate changes in *local* population dynamics with motor learning. However, it remains unclear: (1) how *cross-area* dynamics between M1 and M2 are coordinated and change with learning, and (2) how *local* dynamics in each area might interact with *cross-area* M1-M2 dynamics to drive learning. Previous work on cross-area interactions during motor learning has focused on macroscopic population activity, such as local field potentials [12–16] and wide-field calcium signals [4,17]. However, such measures of aggregate activity inherently collapse signals from a heterogeneous population of neurons into a single measure, making it difficult to resolve potentially important multiplexed signals within that population [18–20].

How then can we distinguish and compare local and cross-area population dynamics during learning? One approach is to use dimensionality reduction methods [21,22] to capture patterns of shared variance within each local population, and then compare those simplified local representations [5,18]. However, since the purpose of dimensionality reduction is to limit the number of signals analyzed, any activity patterns which do not dominate local variance are discarded. Thus, this potentially dismisses as ‘noise’ neural fluctuations that represent activity coordinated across areas. Instead, cross-area activity might be identified by directly detecting covariance which is coordinated across populations. Importantly, recent work in anesthetized animals has shown that simultaneous recordings from visual areas can be analyzed to identify a neural "communication subspace" defined by activity in each region that is maximally correlated with activity in a partner region [23]. However, it is unknown whether such a communication subspace is relevant for behavior and learning.

Thus, this study aims to: (1) measure interactions between *cross-area population dynamics* shared by M2 and M1 with *local population dynamics* compartmentalized to either M2 or M1; and (2) assess the behavioral relevance of *cross-area population dynamics* during motor skill learning. We hypothesized that M2-M1 cross-area dynamics coordinate information between the regions and contribute to learning complex behaviors. We specifically used multisite recordings in M2 and M1, along with dimensionality reduction techniques that capture the multiple axes of variance within and across areas. To capture local dynamics, we used well-known dimensionality reduction methods, which constrain representations of high-dimensional neural activity to axes of maximal local variance (i.e. local subspaces) [4,5,11,19,22,24–33]. To capture cross-area dynamics, we identified communication subspaces between M1 and M2 populations; hereafter, we use the term "cross-area" to refer to activity in each area which is maximally correlated with activity in the partner region (Fig. 1.1). We thus aimed to specifically identify cross-area dynamics and distinguish them from local population dynamics during both early exploratory learning and late learned execution of a skilled movement.

In each region, we found that local and communication subspaces were distinct throughout learning, reflecting separation of local and cross-area population dynamics. Strikingly, not only did cross-area dynamics clearly encode single-trial reaching behavior, these dynamics also became coordinated with local dynamics over learning. Consistent with this functional role, M2 inhibition in well-trained animals impaired reach behavior and disrupted coordination between local and cross-area reach encoding in M1. Together, our results indicate that cross-area M2-M1 population dynamics are important for driving skilled movements, in part through their interaction and influence on local dynamics.

Results

Our model proposes that population activity consists of multiplexed local and cross-area dynamics generated from overlapping sets of neurons (Fig. 1.1a). Thus, each neuron's spiking activity can contribute to both local and cross-area dynamics. To identify population dynamics local to either M2 or M1, we used factor analysis (FA) to find linear combinations of neural activity that maximized shared variance between neurons within the area [7,11,21]. In each region's high-dimensional population activity space, where each dimension corresponds to one neuron's activity, the neuron weights obtained using FA define a 'local subspace' (Fig. 1.1b), representing dominant local signals. In parallel, to identify neural activity coordinated between M2 and M1, we used canonical correlation analysis (CCA) to find linear combinations of M2 and of M1 activity that are maximally correlated with each other. The neuron weights obtained using CCA define a 'communication subspace' (Fig. 1.1b) [23], representing activity that is shared or coordinated between M2 and M1 (see Materials and Methods). Note that M2 and M1 each have both a local subspace (defined by FA) and a communication subspace (defined by CCA). The projections of high-dimensional neural activity onto the local and communication subspace axes provide low-dimensional representations of local and cross-area activity (Fig. 1.1c).

To analyze how functional interactions between local and cross-area M2-M1 neural population dynamics contribute to skill learning, we performed simultaneous recordings of population neural activity in M2 and M1 (Fig. 1.S1) in rats learning a cue-driven reach-to-grasp task, a well-established model for motor skill learning (Fig. 1.2a) [30,34,35]. Importantly, both M2 and M1 are required for learning and performance of reach-to-grasp movements in many model systems, including rodents, non-human primates, and humans [36,37]. Consistent with past studies, animals learned to successfully retrieve pellets with training; there were also concomitant improvements in movement speed and reaction time, which was measured as the time of reach onset relative to the sound cue marking door opening (Fig. 1.2b-d, Fig. 1.S2, quantification in figure legend).

Distinct local and cross-area covariations within neural population activity

We first examined whether M2-M1 cross-area population activity is separable from local population activity. If cross-area coordination is based on locally-defined covariations propagating between brain regions, then we would expect the cross-area activity identified by CCA to be identical to the locally-shared activity identified by FA. In other words, if there is only a single meaningful pattern of covariation both within and across areas, then the subspaces defined by cross-area and local covariations using CCA and FA should be similar. In contrast, we hypothesized that local and cross-area population activity are distinct and that the two methods would therefore identify different subspaces of covariation.

We verified that local and cross-area population activities were distinct in two ways. First, we found that neurons were assigned different weights when constructing local versus cross-area activity subspaces, suggesting that neurons have differential contribution to local computations versus cross-area communication (Fig. 1.2e). Second, we calculated the angle of alignment between local and communication subspaces; we found that local and communication subspaces were distinct, but not orthogonal (Fig. 1.2f, quantification in figure legend; see Materials and Methods). Similar results were obtained using PCA, which captures the total variance in the population activity, rather than the shared variance (Fig. 1.S3). This suggested that some neurons' activity contributes to both local computations and cross-area communication, while other neurons contribute primarily to local or cross-area dynamics.

Local and cross-area covariation patterns remain distinct over learning

We next asked whether the separability of local and cross-area neural activity changed with learning. If learning leads to increased communication between M2 and M1, one potential mechanism is increased alignment of local and communication subspaces. To test this, we calculated the mutual information between neuron weights defining the communication and local neural subspaces and found no significant change with learning (Fig. 1.2e, quantification in figure legend). In other words, a neuron's local subspace weight was no more informative about its communication subspace weight after learning than before (and vice versa). Additionally, the angle between local and communication subspaces did not change with learning (Fig. 1.2f). The consistent alignment of these subspaces demonstrates that the dominant local covariation patterns remain distinct from dominant cross-area covariation patterns. This suggests that activity patterns defining cross-area coordination are distinct from local computations throughout learning.

Correlation of cross-area activity across cortical regions

Since the alignment between local and communication subspaces remains consistent over learning, we next asked whether the correlation of cross-area activity patterns changed with learning. One possibility is that M1 and M2 cross-area activity is less correlated during exploratory behavior and becomes more correlated during skilled behavior, indicating a change in M2 to M1 transmission efficacy. Since CCA finds M1 and M2 communication subspaces with maximal correlation, if M2 to M1 transmission efficacy increases, we would expect the correlation of M1 and M2 CCA-defined subspaces to be higher during skilled behaviors than during early exploratory actions. This would indicate that the M1 and M2 cross-area activity generally becomes more correlated with learning.

To address this, we correlated M1 and M2 cross-area activity during behaviorally relevant time windows (i.e. 2 seconds peri-reach for each trial) during three types of behavior: spontaneous behavior, exploratory reaches in early learning, and directed reaches in late learning (Fig. 1.3a). To our surprise, there was no difference in the mean correlation values (R^2) of M1 versus M2 cross-area activity during different behaviors (Fig. 1.3b, quantification in figure legend). Thus, generally increased coordination between M2 and M1 activity by itself seems unlikely to the drive performance gains observed with learning.

Learning drives encoding of reach initiation in cross-area population dynamics

An intriguing alternative is that learning is due to a change in the task encoding of cross-area signals. Specifically, signals within the existing range of cross-area activity may be remapped to encode information about the task. Thus, while the overall range and correlation of M1-M2 coordinated activity may not change (Fig. 1.3a), high amplitude cross-area activity may now be associated with a particular behavioral state. As noted above, we observed that the door open cue was more rapidly followed by reach initiation after learning (Fig. 1.2d). This suggested that the timing of reach initiation might be an important marker of learning. We thus explored whether M1-M2 cross-area activity could account for this change. To visualize this possibility, we plotted M1 communication subspace activity versus M2 communication subspace activity during the pre-reach period and after reach initiation (Fig. 1.4a). The histograms show the probability density functions of the respective subspace activity before and during the reach. Interestingly, the two behavioral states were significantly more separable after learning (Fig. 1.4b, quantification in figure legend), suggesting that the high amplitude activity coordinated between M1 and M2 gained task relevance with learning.

We hypothesized that this increase in task relevancy allows M2 to trigger reach initiation in M1 through the communication subspace. Consistent with this, peaks in communication subspace activity became associated with reach initiation after learning (Fig. 1.4c). We quantified this association across trials by building a logistic regression model to distinguish communication subspace activity during 2 seconds before reaching versus during reach initiation. Strikingly, detection of reach initiation based on this communication subspace activity model improved with learning (Fig. 1.4d). Using the logistic regression model, we could then probe the time course of reach initiation prediction based on M1-M2 cross-area activity. (Fig. 1.4e, same trials as c). On average, while the time of reach initiation was not well predicted during early trials, it became highly predictable after learning (Fig. 1.4f).

Learning drives encoding of reach efficiency in cross-area population dynamics

Does M2 only send an initiation signal to M1, or instead does M2 input also affect other aspects of the reach? To address this, we examined whether single-trial M2-M1 cross-area population dynamics were informative about single-trial reaching behavior, and whether reach-specific content of transmitted signals increased with learning. Visualizing single-trial activity is essential for behaviors with high variability since trial-averaging is likely to obscure behavior-related signals. To quantify reach modulation in single-trial neural activity, we calculated a communication subspace neural modulation metric (CS modulation), which compares neural activity during reaching versus an equivalent baseline period for each trial (Fig. 1.5a,b). This measure is equivalent to the d' ('d-prime') signal sensitivity index used in signal processing (see Materials and Methods). To directly test the relationship between behavioral performance and the M1 and M2 CS modulation, we correlated neural CS modulation with reach duration on a trial-by-trial basis (Fig. 1.5c, quantification in figure legend). Interestingly, we found that CS modulation reliably predicted reach duration, indicating that cross-area population dynamics encoded additional behaviorally relevant information. Additionally, both M1 and M2 CS modulation increased with learning (Fig. 1.5d). Thus, the process of learning appeared to enhance reach-specific neural signals in cross-area population dynamics, providing a mechanism for coordinating network-wide activity related to tasks being learned.

Learning drives coordinated encoding of reach efficiency in local and communication subspaces

Cross-area population dynamics may provide a mechanism for coordinating network-wide task activity. Indeed, our overarching model of learning proposes that reaching signals become coordinated between cross-area and local dynamics to drive learning. Specifically, we expected local and cross-area neural reach modulation to become coordinated with learning (Fig. 1.6a). This would support a learning model in which M2 local dynamics develop task-specific activity which is then transmitted through M2-M1 cross-area dynamics to M1 local dynamics (Fig. 1.6b). We directly measured concordance of trial-to-trial reach modulation between local and cross-area signals in each region (Fig. 1.6c-f). Strikingly, we found that normalized mutual information between the reach modulation of local and cross-area signals in M2 and M1 increased with learning (see Materials and Methods; Fig. 1.6d,f, quantification in figure legend). Specifically, trials in which the cross-area dynamics were highly modulated by movement also tended to have high neural modulation in the local dynamics, indicating increased coordination between local computations and transmitted signals related to reaching.

M2 inactivation disrupts skilled reaching

A prominent model of M2-M1 interactions during learning proposes a strong top-down influence from M2 to M1[3,4]. If activity transmitted from M2 to M1 drives M1 reach encoding, then disrupting M2 to M1 transmission would impact reaching behavior. To test this, we inactivated M2 in well-trained animals using the GABA agonist muscimol (Fig. 1.7a,b). Unlike control saline infusions (Fig. 1.S4), M2 inactivation caused performance deficits, with reaching behavior qualitatively similar to early learning (Fig. 1.7a,c; Fig. 1.8a, quantification in figure legend). However, the mechanism of this deficit is unknown. We hypothesized that M2 inactivation disrupts M1-M2 cross-area population dynamics, thereby removing top-down influence on M1 local dynamics without disrupting local connectivity.

M2 inactivation disrupts M1 encoding of reach initiation

We next performed simultaneous recordings in M1 and M2 during baseline performance and during M2 inactivation on the same day, in well-trained animals. This approach allowed us to track the effect of M2 disruption on M1 cross-area and local dynamics (Fig. 1.7d). First, we found that M2 inactivation disrupted encoding of reach initiation in the M1 communication subspace (Fig. 1.7d-f). We quantified this by comparing the difference in median activity before reach and at reach initiation (Fig. 1.7e, quantification in figure legend), and found that this difference was significantly smaller during M2 inhibition. As before, we fit a logistic regression model to predict reach onset from M1 communication subspace activity. We quantified the model's performance and saw that detection of reach initiation based on M1 communication subspace activity decreased with M2 inhibition (Fig. 1.7f), indicating that M1 cross-area dynamics were less informative about reach initiation during M2 inhibition.

M2 inactivation disrupts M1 encoding of reach efficiency

In addition to disrupting reach initiation signals, we also found that M2 inhibition disrupted reach modulation of M1 cross-area and local dynamics (Fig. 1.8d, f, quantification in figure legend). This indicated that M2 input is necessary for intact M1 reach modulation and implied a M2 to M1 directionality. We additionally examined whether M2 inactivation entirely dissociated M1 CS reach modulation from behavioral performance. We found that the relationship between reach duration and M1 CS modulation was still significant during M2 inactivation (Fig. 1.8e), underscoring the fundamental relationship between M1 and behavior. Finally, we tested whether M2 inactivation disrupted coupling between M1 cross-area and local dynamics. Interestingly, we found that the mutual information between the single-trial reach modulation of M1 local and communication subspaces decreased significantly with M2 inactivation (Fig. 1.8g), indicating a decoupling between the local and cross-area dynamics. This decoupling may provide a mechanism for resilience of local dynamics, which could create robustness in the event of distant network damage. Importantly, the changes in M1 cross-area dynamics were not due to changes in overall M1 firing rate, which did not change significantly (23.4 Hz ± 4.0 for Baseline; 18.6 Hz ± 3.3 for M2 Muscimol; mixed effect model, p = 0.157). Furthermore, mean M1 local covariance did not change, indicating stability in local M1 connectivity (0.24 ± 0.06 shared variance/total variance for Baseline; 0.19 ± 0.04 shared variance/total variance for M2 Muscimol; mixed effect model, p = 0.458, see Materials and Methods).

Discussion

This study outlines a new approach to understanding interactions between two nodes in a neural network. We analyzed how dynamics in a cross-area communication subspace interact with local population dynamics. First, we showed that computational methods that maximize either local or cross-area covariance identify distinct local and communication subspaces. This suggests that activity that might have been previously considered 'noise' by local-only dimensionality reduction methods may actually contain important signals transmitted from a partner area. Second, we show that cross-area population dynamics become markedly more related to both reach initiation and reach duration with learning. Through causal manipulations, we found that local M2 inactivation disrupted M1 cross-area population dynamics as well as reach execution. The remnant M1 cross-area population dynamics were attenuated but still predictive of single-trial behavior, indicating maintenance of meaningful activity in M1. However, the attenuation of M2's influence on M1 local population dynamics prevented top-down guidance of learned behavior, i.e. slower reaction to environmental cues and less efficient reaches. These results demonstrate that communication and local subspaces are distinct, that learning shapes the content of their shared information, and that execution of learned skills depends on transmission of top-down task information through cross-area population dynamics.

Distinct local and cross-area population dynamics enable flexible communication

A key result of our study is that methods that maximize local variance, such as FA and PCA, find different population dynamics than methods that maximize covariance between regions, such as CCA (used here) or reduced rank regression (used in ref. 23). Although activity projected onto local and communication subspaces may look qualitatively similar in well-trained animals (Fig. 1.6a), neuron weights remain distinct and the subspaces have a consistent ~ 45º angle between them throughout learning (Fig. 1.2e,f), indicating that the computations underlying these dynamics remain distinct. Segregation of local computations and communication processes could allow for selective information routing [23]. For example, M2 may share a different communication subspace with striatum [38], allowing M2 to send different signals to M1 and striatum. A similar separation between local and communication subspaces has been found in visual cortex[23]. As distinct subspaces for local and cross-area population dynamics have now been identified in both sensory and motor systems, functional compartmentalization may be an important general principle of communication between nodes of cortical networks. By showing that cross-area population dynamics can explain both learning gains and behavioral deficits resulting from M2 inactivation, our work provides evidence that such communication subspaces have functional, behavioral relevance.

The idea that cortical regions may communicate via patterns of coordinated population dynamics presents an alternative understanding of functional connectivity to well-known theories like communication through coherence [39]. While communication through coherence relies on gating communication through phase alignment, communication through subspaces relies on routing information through functional cross-area dynamics, irrespective of downstream activity or the presence of oscillations. The presence of neurons with high weights for both local and communication subspaces suggests a mechanism for calibrating the strength of information transmission between cross-area and local computations. Namely, modifying the strength of synaptic connections between the dominant cross-area neurons and local neurons with weaker cross-area connections – in either the upstream or downstream region – could change the strength of signal transmission, allowing for bi-directional control over communication.

Cross-area population dynamics explain single-trial behavior

Using computational methods that specifically identified cross-area population dynamics, we found that learning differentially affects the *correlation* of M2-M1 cross-area activity and the *mapping* of cross-area population dynamics relative to behavior. The maximum correlation strength between M2 and M1 cross-area population dynamics did not change with behavior (Fig. 1.3), indicating that the connectivity that determines shared variance between the two regions is not determined by behavioral states. However, learning did strengthen the coordination of task information between local and cross-area population dynamics (Fig. 1.6), as well as their link to behavior (Fig. 1.4, 1.5). Our results suggest that the cross-area dynamics are an important mediator of this change. This is further supported by observed lack of a change in the correlations in communication space between the two areas with learning.

Although learning was not associated with changes in the correlation strength of cross-area population dynamics, it was associated with changes in their *mapping* to behavior. Past work has proposed that the role of M2 is to provide top-down control and contextual information to M1 [3–5,17,40]. Here, we provide insight into what such a signal might look like, and how it evolves with learning. In early learning, when behavior was exploratory and highly variable, high amplitude cross-area dynamics were less related to specific behavioral timepoints (Fig. 1.4a,c), and modulation of communication subspace activity was only weakly related to reaching (Fig. 1.5a,d). However, even at this early stage, reaches with higher communication subspace modulation tended to have shorter durations (Fig. 1.5c). After the task was learned, the relationship between communication subspace modulation and behavior was amplified (Fig. 1.4, 1.5). Notably, the single-trial M2-M1 cross-area dynamics corresponding to similarly efficient, short duration reaches in early and late learning were not identical in early and late learning (Fig. 1.5c). This argues against the notion that pre-existing representations of efficient movements are simply selected for through the process of learning. Instead, our results support the idea that learning transforms [41] and amplifies the neural signals for behaviors that are being selected. This finding also highlights the feasibility and importance of analyzing single-trial neural activity and behavior in order to understand highly variable behavioral states such as early learning.

"On-manifold" causal manipulation of downstream neural activity

Finally, the relationship between cross-area population dynamics and behavior appears to be causal, since M2 inactivation disrupted both M1 cross-area population dynamics and reaching behavior (Fig. 1.7, 1.8) while leaving local properties of M1 intact (i.e. firing rates and proportion of variance shared locally). Examining local activity during upstream inactivation provides a valuable approach to differentiating between activity dynamics generated locally and those propagated from top-down influences. Such analyses are impossible in purely correlative studies, and, paired with same-day establishment of cross-area dynamics, demonstrate a novel approach to understanding how several axes of variance and information encoding overlap [42] and interact within functional neural systems. Furthermore, our simultaneous recording of M1 activity during M2 inactivation demonstrates that, when M2 inputs are removed, M1 is still within its native manifold, as measured through M1 mean firing rate and local shared variance. This is important because there has been increasing concern that acute changes in input to an area can perturb behaviorally relevant local population dynamics [43,44]. Importantly, rats do produce some successful reaches both during M2 inactivation and in early learning, although they are more infrequent and less efficient than during the intact learned state. Together, this demonstrates that M1 is independently capable of producing functional reach-to-grasp behavior, and that the top-down input from M2 is a learned signal, biasing M1 towards more effective behavior. This is concordant with long-standing models of top-down M2-M1 interactions during learning [3] and reinforces the view that, while M2 and M1 both contain representations of movement, M2 is particularly important for learned, complex skills [2,4,45,46].

Conclusion

Our results provide direct evidence that M2-M1 cross-area neural population dynamics, that are increasingly modulated by task learning and performance, become coupled to local population dynamics in M2 and M1 with learning. Knowledge of this phenomenon should help to better understand mechanisms of neural plasticity and functional properties of large-scale, hierarchical networks in the context of flexible, learned skilled motor behaviors.

Materials and Methods

Animal Care.

All procedures were in accordance with protocols approved by the Institutional Animal Care and Use Committee at the San Francisco Veterans Affairs Medical Center. Adult male Long Evans rats (*n* = 10, 250–400 g; Charles River Laboratories) were housed in a 12-h/12-h light–dark cycle. All experiments were done during the light cycles. Rats were housed in groups of 2 animals prior to surgery and individually after surgery.

Surgery.

All surgical procedures were performed using a sterile technique under 2–4% isoflurane. Surgery involved cleaning and exposure of the skull, preparation of the skull surface (using cyanoacrylate) and then implantation of the skull screws for overall headstage stability. Reference screws were implanted posterior to lambda and ipsilateral to the neural recordings. For experiments involving physiological recordings, craniotomy and durectomy were performed, followed by implantation of the neural probes. For experiments involving only infusions, burr holes were drilled in the appropriate locations, followed by implantation of the cannulas. Postoperative recovery regimen included the administration of 0.02 mg per kg body weight buprenorphine for 2 days, and 0.2 mg per kg body weight meloxicam, 0.5 mg per kg body weight dexamethasone and 15 mg per kg body weight trimethoprim sulfadiazine for 5 days. All animals were allowed to recover for 1 week prior to further behavioral training.

Electrode array and cannula implants.

Rats were implanted with two 32-channel tungsten wire probes (TDT or Innovative

Neurophysiology), one each in M1 (+0.5 AP, +3.5 ML, -1.5 DV) and M2 (+4.0 AP, +1.5 ML, -1.5 DV), contralateral to reaching arm. Infusion cannulas were implanted in M2 (+4.0 AP, +1.5 ML, -1.5 DV) for infusion-only animals. For rats with both M2 electrode arrays and cannulas, the cannula was attached to the electrode array prior to surgery.

Pharmacological infusions.

Rats were anesthetized with 2% isoflurane before infusions. We injected 0.5 - 1uL (1 µg/µl) [47] of the GABA receptor agonist muscimol into contralateral M2 (infusion rate: 1nl/min) through a chronically implanted cannula using a Hamilton infusion syringe. The infusion syringe was left in place for at least 5 min post-infusion. Rats were allowed to recover in their home cages for 2 hours before starting behavioral testing.

Histology.

Final placement of the electrodes was monitored online based on implantation depth and verified histologically at the end of the experiments. Rats were anesthetized with isoflurane and transcardially perfused with 0.9% sodium chloride, followed by 4% formaldehyde. The harvested brains were post-fixed for 24 h and immersed in 20% sucrose for 2 days. Coronal cryostat sections (40-µm thickness) were mounted with permount solution (Fisher Scientific) on superfrosted coated slides (Fisher Scientific). Images of a whole section were taken by a HP scanner, and microscope images were taken by a Zeiss microscope.

Behavioral training.

We used an automated behavior paradigm to train rats to perform dexterous reach-to-grasp movements[30,35]. Rats learn to reach through a narrow slot to grasp and retrieve a 45 mg pellet from a shallow dish (i.e. pellet holder) placed ~1.5 cm outside the behavioral box [34]. Prior to implantation, rats were handled and habituated to the behavioral box for at least one day, then manually prompted to reach for a pellet 10-30 times to determine handedness. Handedness was determined when rats reached with the same hand for >=70% of at least 10 test trials. The start of each trial was signaled with a tone and the opening of a door allowing access to the pellet. Trials ended when the door was closed, which was triggered either by the pellet being dislodged from the pellet holder, or, if this did not occur, ~15s after door opening.

Behavioral training for learning animals.

Once handedness was determined, rats were implanted with neural probes (see Surgery). For two days before behavioral training, rats were food restricted, followed by feeding animals a fixed amount during the course of training. During behavioral training, rats were placed in an automated reach box and completed 38-300 trials per day. The ‘early learning’ training day was the first day on which the rat completed at least 30 trials. The ‘late learning’ training day was the second consecutive day on which the rat performed with at least 45% success rate.

Behavioral training for M2 inactivation animals.

Once handedness was determined, rats were trained until their success rate reached a plateau (>2 consecutive days with performance above 45% and > 100 completed trials/day), after which they were implanted with infusion cannulas alone (n=3 rats), or with infusion cannulas and electrodes (n=3 rats) (see Surgery). Rats were allowed at least a week of recovery after surgery before beginning behavioral testing. Rats were re-trained until plateau performance (>2 consecutive days with performance above 40%). On M2 inactivation days, rats performed 100 reach trials before receiving pharmacological infusions. After 2 hours of rest post-infusion, rats were re-tested for 100 trials.

Behavioral analysis.

Rat behavior was video recorded using a side view camera (30 - 100 Hz) positioned outside the behavioral box, perpendicular to the main direction of movement. Each rat's reach hand was painted with an orange marker at the start of each day. Reach videos were viewed and semi-automatically scored to obtain trial success, hand position, and time points for reach onset, and grasp onset. To characterize motor performance, we quantified reach duration, distance travelled, maximum movement speed, and pellet retrieval success for each trial. Percent reach success is the percent of trials on which the pellet was retrieved during a single day of training, excluding trials in which the rat did not dislodge the pellet from the holder or displayed abnormal behavior (i.e. licking, reaching with the wrong hand). Reach duration for each trial was defined as the time from the start of reach to onset of grasping or when the paw first touched the pellet if no grasping occurred on that trial. Reach distance was the sum of the path travelled during that time.

Electrophysiology data collection.

We recorded extracellular neural activity using tungsten microwire electrode arrays (MEAs, n = 7 rats, TDT or Innovative Neurophysiology). We recorded spike and LFP activity using a 128–channel TDT–RZ2 system (TDT). Spike data was sampled at 24,414 Hz and LFP data at 1,018 Hz. Analog headstages with a unity gain and high impedance (~1 GΩ) were used. Snippets of data that crossed a high signal-to noise threshold (4 standard deviations away from the mean) were time-stamped as events, and waveforms for each event were peak aligned. For 2 animals, MEA recordings were sorted offline using superparamegnetic clustering program (WaveClus [48]). For 5 animals, MEA recordings were sorted offline using a density-based clustering algorithm (Mountainsort [49]). Clusters interpreted to be noise were discarded, but multi-units were kept for analysis. Trial-related timestamps (i.e., trial onset, trial completion, removal of pellet from pellet holder, and timing of video frames) were sent to the RZ2 analog input channel using an Arduino digital board and synchronized to neural data.

Neural data analysis: local neural subspace and population dynamics.

We used Factor Analysis (FA) to define local neural dynamics [50,51]. FA models the joint distribution of N neurons' spike counts (rank N) as the sum of a mean rate *d* for each neuron (rank N), private signals with diagonal covariance R (rank N x N), and shared signals corresponding to latent factors *z* (rank k, k < N).

To estimate the number of latent dimensions in each dataset, we performed 5-fold cross-validated FA on the dataset using k = 1:N factors and estimated the log likelihood from each iteration. We averaged the log-likelihood from the 5 iterations for each candidate dimensionality and identified the dimensionality which yielded the highest log likelihood. We then fit using this dimensionality and estimated the number of dimensions needed to account for 75% of the shared variance, elsewhere referred to as the 'main shared variance' [7]. The mean value across all datasets was 3.7, and we conservatively chose to use k = 3 factors for all of our analyses (for all datasets 3 < N).

To visualize the time course of shared variance on each trial, we used FA to create neural trajectories of each region's population firing on each trial. The models were built using neural data binned at 100ms, from -1s to +1s surrounding the time of grasp onset, concatenated for all trials. Results were not qualitatively different if only data from reach onset to grasp onset was included to build the model. For visualization only, data was interpolated to 10 ms resolution using a spline fit.

Neural data analysis: cross-area neural subspace and population dynamics.

Communication subspaces were defined using Canonical Correlation Analysis (CCA), which identifies maximally correlated linear combinations between two groups of variables [52]. Neural data in M2 and M1 was binned at 100ms, and data from -1s to +1s surrounding time of grasp onset was concatenated across trials. CCA models were fit using the MATLAB function canoncorr. The models' performance was evaluated using the R^2 of the top canonical variable (CV) across 10-fold crossvalidation. Significant predictive performance was calculated by comparing the R^2 of each canonical variable to the R^2 of the top CV in a trial-shuffled bootstrap distribution. Some datasets had 2 or 3 significant CVs, but we worked with the top CV only since the top CV was significant for all datasets. Results were not qualitatively different if only data from reach onset to grasp onset was included to build the model. For visualization only, data was interpolated to 10 ms resolution using a spline fit.

Neural data analysis: subspace alignment.

The alignment between the subspaces defining the local and communication subspaces was calculated using the MATLAB function ‘subspace’. Weights for all 3 factors were included for the local subspace. Weights for only the top 1 canonical variable were included for the communication subspace.

Neural data analysis: reach start signaling

To calculate the difference in communication subspace (CS) activity before reach initiation versus during reach initiation, we defined a ‘pre-reach period’ as -2s to -0.1s before reach initiation and a ‘reach initiation’ period from -0.1s to +0.3s surrounding reach initiation. CS activity from each of these periods was concatenated across trials to then calculate the median CS activity value. The difference between median CS activity during pre-reach and reach initiation was calculated for each animal. Statistics were calculated using mixed effect modeling across animals.

For reach start prediction, activity from pre-reach and reach initiation was labelled as ‘0’ or ‘1’, respectively, which was then used as the response values to train a logistic regression model using the MATLAB function ‘fitglm’. The probability that CS activity values corresponded to a timepoint during reach initiation was returned as scores. We then used these scores to compute the receiver operating characteristic (ROC) curve of the classification results using the MATLAB function ‘perfcurve’. The area under the curve (AUC) was returned for each animal, and these values were used in mixed effect modeling to detect difference in pre-reach versus reach initiation activity during early versus late learning, and baseline versus muscimol behavior.

The logistic regression model was used to calculate the probability of reach initiation based on CS activity on single trials. We calculate the single-trial difference in the mean predicted probability of reach initiation during the pre-reach versus reach initiation periods. We compared this difference using all trials in early versus late learning, and baseline versus muscimol behavior. We plotted the median probability of reach initiation across trials aligned to reach initiation.

Neural data analysis: neural reach modulation.

Single-trial neural reach modulation of each factor defined using FA and each canonical variate defined using CCA was calculated using the signal processing d’ (d-prime) signal sensitivity metric defined by the equation below [53], where μ indicates the mean and σ indicates the standard deviation of the signal. For each trial, the ‘reach’ period was defined as -0.1 s before reach onset to + 0.1s after grasp onset; the ‘baseline’ period was defined as a length of time equal to the reach period, ending 1s before the start of the reach period. For each trial, the signal was the absolute value of the difference between each datapoint and the mean of the baseline period. The median value from the baseline period was subtracted from both the movement signal and the baseline signal before calculating the single-trial modulation value (d’), as below. For Fig. 1.6, the calculation of mutual information between local and cross-area signals, both signals were normalized to their max values before calculating the single-trial modulation value (d’).

$$d' = \frac{\mu_{reach} - \mu_{baseline}}{\frac{1}{2}\sqrt{\sigma_{reach} + \sigma}_{baseline}}$$

Each factor defined using FA defined activity of one local subspace axis. Activity of all factors was included in calculations of overall neural reach modulation. When activity of only one factor is visualized, we chose the factor accounting for the largest proportion of shared variance (i.e. the top factor).

Neural data analysis: mutual information.

To calculate the concordance in neural reach modulation in trial-to-trial local and cross-area neural reach modulation, we used the mutual information equation below. X and Y are the set of single-trial neural reach modulation values from local and communication subspace activity.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) * log \frac{p(x,y)}{p(x)p(y)} dx\, dy$$

To obtain the normalized mutual information, this value was divided by $\sqrt{p(x)p(y).}$

Neural data analysis: mean local covariance.

Each neuron's shared over total variance was calculated as in Athalye et al., 2017 [7]. Briefly, the subspace of shared variance is represented as the matrix of factor weights U (N x z), where each column contains the weight of each neuron's firing rate for that factor. The covariance matrix is calculated as U^*U^T. Each neuron's variance can be broken down into private variance (the diagonal of R) and shared variance (the diagonal of the covariance matrix). Each neuron's shared over total variance is calculated as shared / (shared + private variance).

Statistical analysis.

Unless stated otherwise, all statistical tests were done using hierarchical mixed-effect models using the MATLAB function 'fitlme' and are written as mean ± SEM. Rat identity was always considered a random effect. When calculating changes in neural reach modulation between early and late learning, we included reach duration as a covariate to control for changes in reach duration between early and late learning. When calculating the relationship between neural reach modulation and reach duration, we included learning stage (early vs. late) as a covariate.

Acknowledgements

We thank C. Campillo Rodriguez for assistance with histology; S.J. Won for assistance with perfusions and imaging; N. Hoglen, L. Tian, P. Khanna, and P. Shirvalkar for providing comments on the manuscript. **Funding:** This work was supported by fellowship awards from the National Defense Science and Engineering Graduate Fellowship (NDSEG, https://ndseg.asee.org/) the UCSF Discovery Fellows Program, and the Markowski-Leach Fellowship, the UCSF Medical Scientist Training Program (to T.L.V.), and the UCSF Neuroscience Graduate Program (to T.L.V. and K.D.). Additional funds come from the Department of Veterans Affairs, Veterans Health Administration (VA Merit: 1I01RX001640 to K.G) and National Institute of Mental Health, NIH (5R01MH111871 to K.G.); and start-up funds from the UCSF Department of Neurology to K.G.. K.G. also holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund (1009855) and an Independent Scientist Award (1K02NS093014) from the National Institute of Neurological Disorders and Stroke, NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. **Author contributions:** T.L.V., K.D. and K.G. designed the study. T.L.V. and K.D. conducted all experiments and analyses. T.L.V., K.D., and K.G. wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The data sets generated and analyzed in the current study (and the associated custom code created in MATLAB) are available from the corresponding author on reasonable request.

Figures

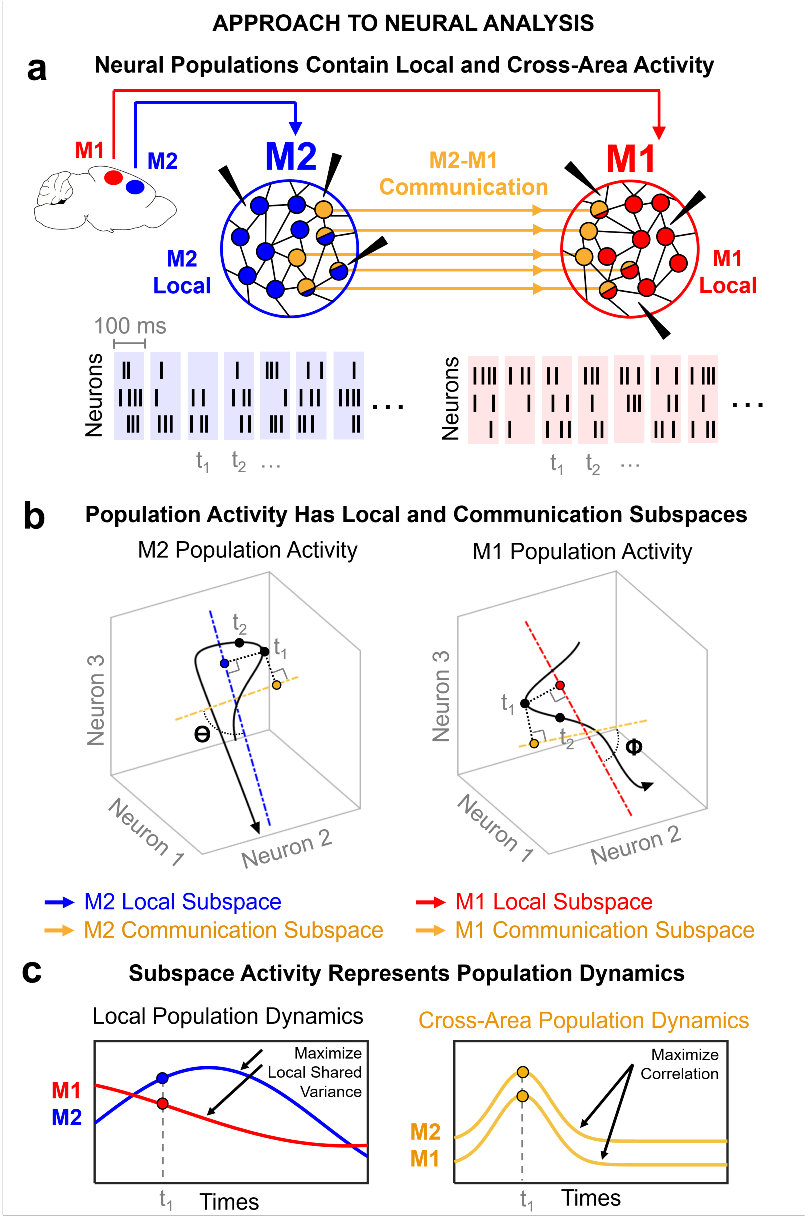

APPROACH TO NEURAL ANALYSIS
a
Neural Populations Contain Local and Cross-Area Activity
M1
M2
M2
M2-M1
Communication
M1
M2
Local
M1
Local
100 ms
Neurons
t1 t2 ...
Neurons
t1 t2 ...
b
Population Activity Has Local and Communication Subspaces
M2 Population Activity
M1 Population Activity
Neuron 3
Neuron 1
Neuron 2
t2
t1
Θ
Neuron 3
Neuron 1
Neuron 2
t1
t2
Φ
M2 Local Subspace
M2 Communication Subspace
M1 Local Subspace
M1 Communication Subspace
c
Subspace Activity Represents Population Dynamics
Local Population Dynamics
Cross-Area Population Dynamics
Maximize
Local Shared
Variance
M1
M2
t1
Times
Maximize
Correlation
M2
M1
t1
Times

Figure. 1.1. Parsing local and cross-area neural signals.

(**a**) Local and cross-area inputs drive neural population activity. (Top) Illustration of neural data being recorded simultaneously in M2 and M1. (Bottom) Activity from each neuron was binned at 100ms. (**b**) Population activity has local and communication subspaces. (Left) A multi-dimensional neural space can be defined using the activity of each M2 neuron as one dimension. Neural population activity was decomposed into local and cross-area signals (dotted lines represent axes in the high-dimensional space). Factor analysis (FA, shown in blue for M2 and red for M1 throughout) was used to uncover signals that were local within M2 or M1. Canonical correlation analysis (CCA, shown in gold throughout) was used to uncover signals that were maximally correlated between M2 and M1. **ϴ** represents the angle between the M2 local subspace defined using FA and the M2 communication subspace defined using CCA. (Right). Same as left, but for M1. **Φ** represents the angle between the M1 local subspace and the M1 communication subspace. (**c**) Subspace activity represents neural population dynamics. (Left) Projections of high-dimensional neural activity on local subspace axes provides low-dimensional readouts of local population dynamics. (Right) Same as left, but for M2-M1 cross-area dynamics.

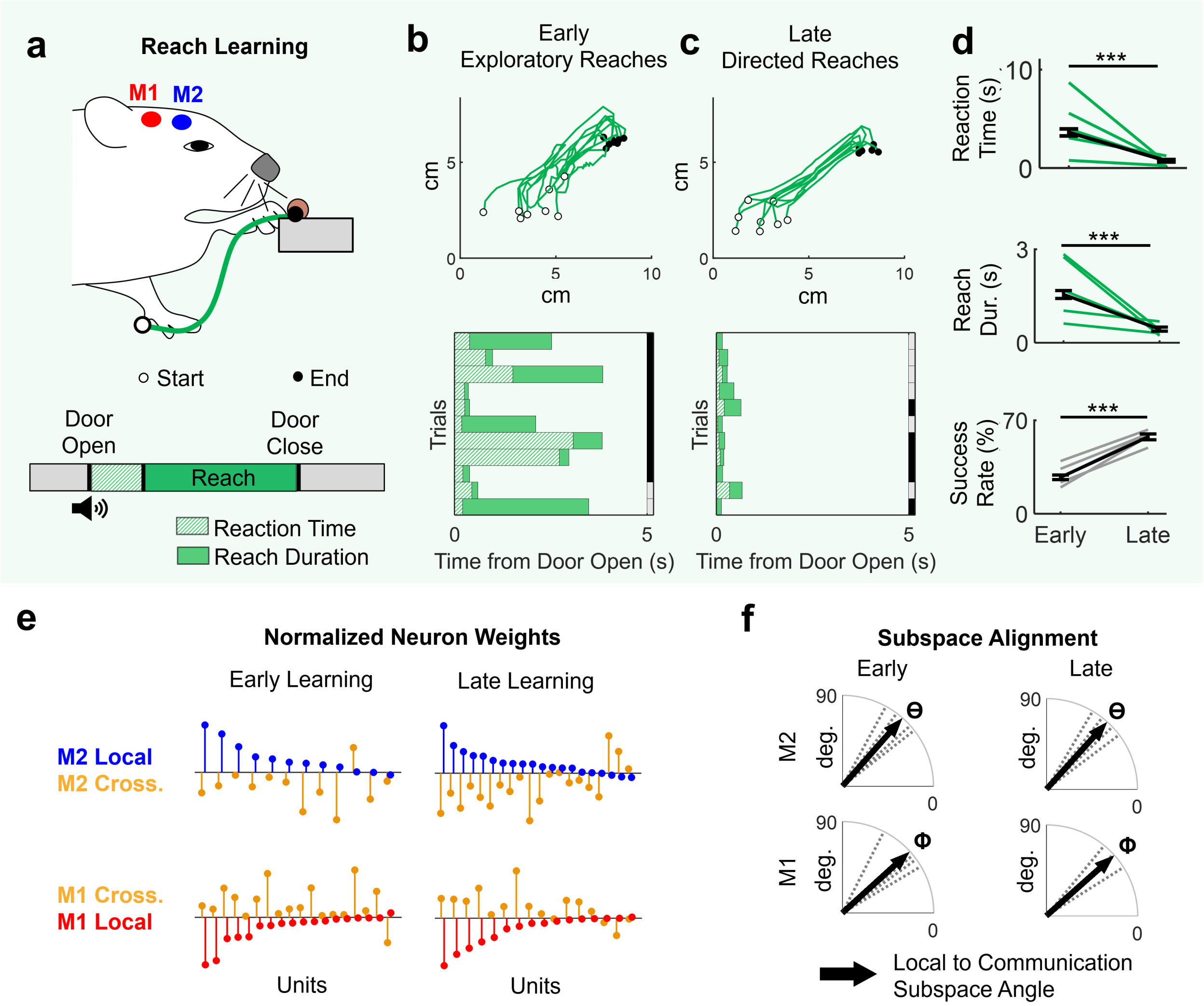

a
Reach Learning
M1
M2
○ Start
● End
Door Open
Door Close
Reach
▨ Reaction Time
🟩 Reach Duration
b
Early Exploratory Reaches
cm
5
0
0
5
10
cm
Trials
0
5
Time from Door Open (s)
c
Late Directed Reaches
cm
5
0
0
5
10
cm
Trials
0
5
Time from Door Open (s)
d
Reaction Time (s)
10
0
***
Reach Dur. (s)
3
0
***
Success Rate (%)
70
0
***
Early
Late
e
Normalized Neuron Weights
Early Learning
Late Learning
M2 Local
M2 Cross.
M1 Cross.
M1 Local
Units
Units
f
Subspace Alignment
Early
Late
M2
M1
deg.
90
0
Θ
Φ
Local to Communication Subspace Angle

Figure. 1.2. Motor behavior engages distinct local and cross-area activity patterns. (**a**) (Top) Rats were trained to perform the reach-to-grasp task. (Bottom) Single-trial experimental paradigm. (**b**) Example reaches in early learning. (Top) Paw trajectories. (Bottom) Example consecutive single-trial representations of reaction time and reach duration. Right border of plot shows accuracy, with success in gray and failure in black. (**c**) As in (b) but for late learning. (**d**) With learning, reaction times decreased (mixed effect model, 3.61s Hz ± 0.36 for Early; 0.73s ± 0.12 for Late, $p = 1.68 \times 10^{-105}$), reach durations decreased (mixed effect model, 1.55s Hz ± 0.12 for Early; 0.43s ± 0.07 for Late, $p = 2.17 \times 10^{-60}$), and success rates increased (mixed effect model, 27.3% ± 1.7 for Early; 57.6% ± 2.1 for Late, $p = 2.75 \times 10^{-42}$). (**e**) Length of stems indicate weights for each neuron's contribution to local or cross-area activity, derived using FA and CCA respectively. Neuron weights were normalized by the maximum value for any neuron for that subspace. FA and CCA weights are shown offset and opposing for visual clarity; for each subspace, most neurons had positive weights. M2 and M1 neuron weights in (Left) early learning and (Right) late learning. Mutual information between FA and CCA weights did not change with learning (mixed effect model, M2: 0.9 ± 0.1 for Early; 0.9 ± 0.1 for Late, p = 0.99, M1: 0.9 ± 0.0 for Early; 0.8 ± 0.1 for Late, p = 0.36). (**f**) Angle in multi-dimensional space between local and cross-area subspaces. Black arrow is the mean angle across animals, dashed lines show values for each animal. In order, M2 (**ϴ**) and M1 (**Φ**) subspace angles in (Left) early learning and (Right) late learning. In each region, the angles between local and cross-area activity axes were significantly different from zero (mixed effect model, M2: $p = 6.11 \times 10^{-5}$, M1: $p = 6.34 \times 10^{-5}$), and are not significantly different between early and late learning (mixed effect model, M2: 44.21 deg ± 4.46 for Early; 45.00 deg ± 6.04 for Late, p = 0.90, M1: 43.51 deg ± 4.75 for Early; 47.69 ± 6.33 for Late, p = 0.54).

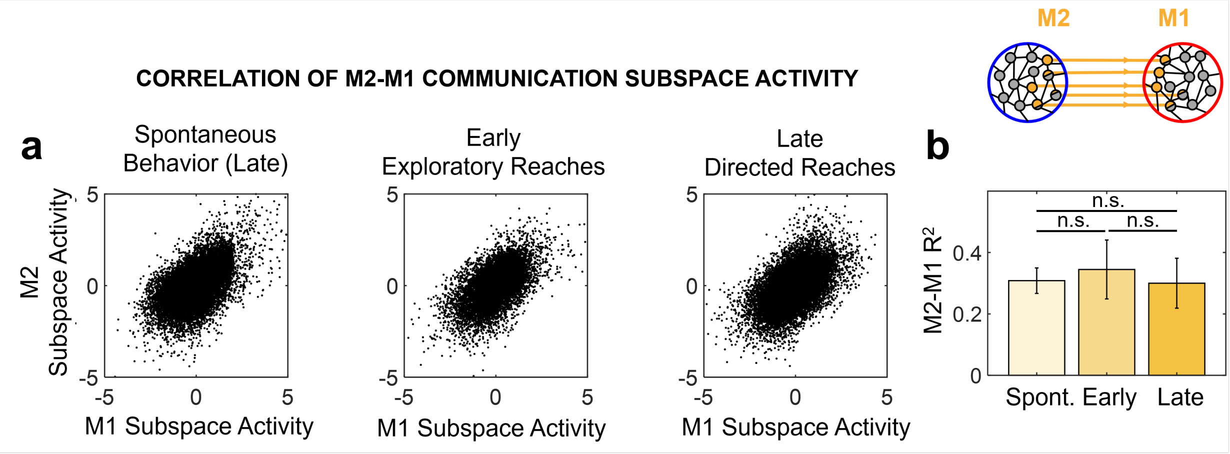


Figure. 1.3. Correlation of M2-M1 cross-area population activity does not change with learning.

(**a**) Correlation between M2 and M1 components of the M2-M1 cross-area population activity during (Left) spontaneous behavior, (Middle) early exploratory reaches, and (Right) late directed reaches. Spontaneous behavior was during the late learning day. Each data point is M2 and M1 data from a single 100ms bin (n = 4 rats). (**b**) Quantification of (A) as correlation R^2 values. Correlation is not significantly different during spontaneous behavior, early reaches, and late reaches (mixed effect model, r = 4 rats; 0.31 ± 0.04 for Spontaneous, 0.34 ± 0.10 for Early, 0.30 ± 0.08 for Early; Spontaneous vs. Early: p = 0.66; Spontaneous vs. Late: p = 0.89; Early vs. Late: p = 0.49).

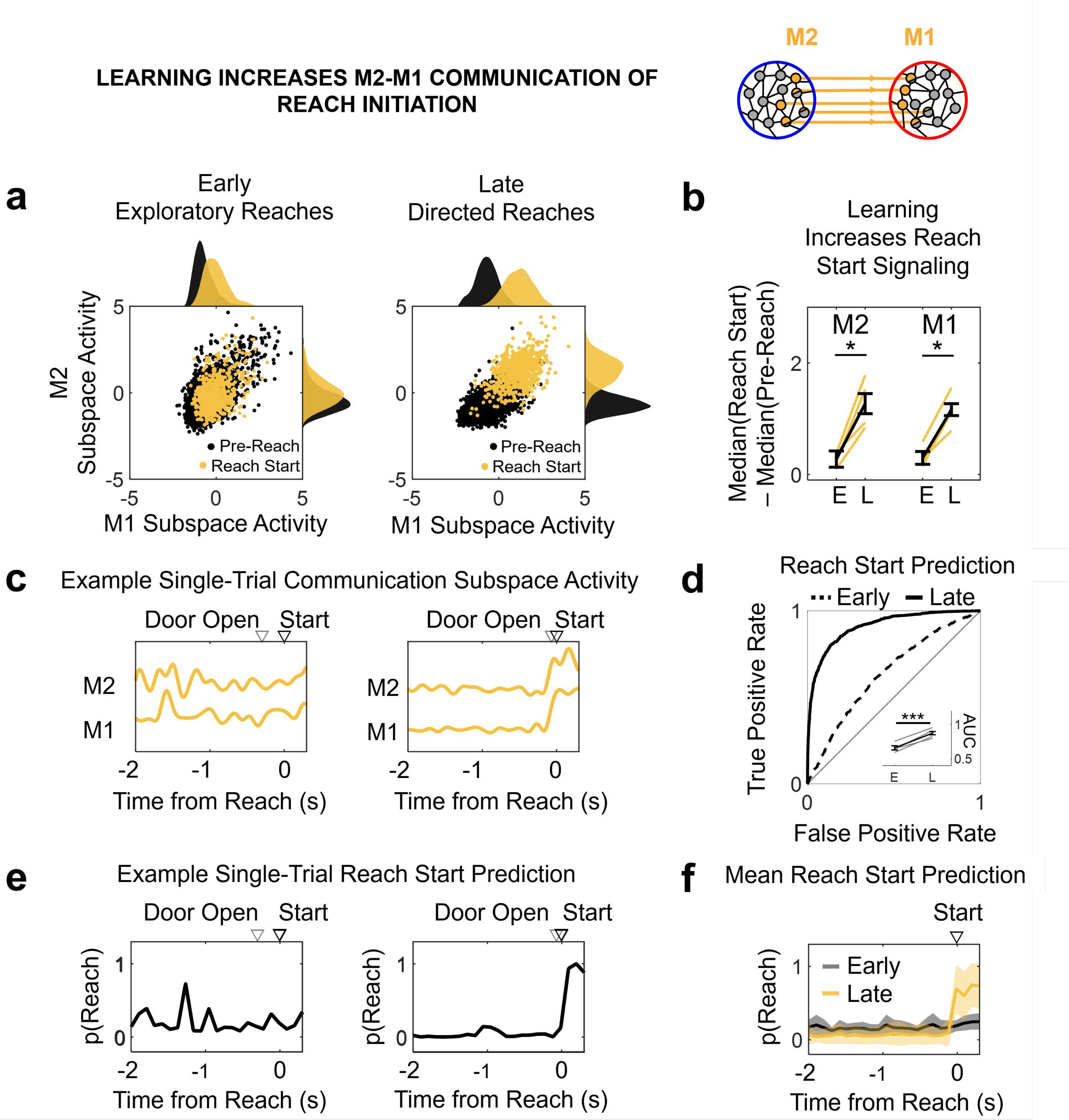

LEARNING INCREASES M2-M1 COMMUNICATION OF REACH INITIATION
M2
M1
a
Early
Exploratory Reaches
Late
Directed Reaches
M2
Subspace Activity
5
0
-5
-5
0
5
M1 Subspace Activity
● Pre-Reach
● Reach Start
b
Learning
Increases Reach
Start Signaling
Median(Reach Start)
– Median(Pre-Reach)
M2
M1
*
*
2
0
E L
E L
c
Example Single-Trial Communication Subspace Activity
Door Open
Start
M2
M1
-2
-1
0
Time from Reach (s)
d
Reach Start Prediction
Early
Late
True Positive Rate
False Positive Rate
1
0
0
1
***
AUC
1
0.5
E
L
e
Example Single-Trial Reach Start Prediction
Door Open
Start
p(Reach)
1
0
-2
-1
0
Time from Reach (s)
f
Mean Reach Start Prediction
Start
Early
Late
p(Reach)
1
0
-2
-1
0
Time from Reach (s)

Figure. 1.4. Learning drives communication subspace encoding of reach initiation.

(**a**) M2-M1 communication subspace activity before reach and during reach initiation for example animal. Probability density functions of M1 (top) and M2 (right). (**b**) Quantification of (a) as the difference between pre-reach and reach median activity during early and late learning for (left) M2 and (right) M1 communication subspace activity (mixed effect model, r = 4 rats. M2: 0.31 ± 0.15 for Early; 1.29 ± 0.18 for Late, p = 0.0017; M1: 0.27 ± 0.14 for Early; 1.09 ± 0.12 for Late, p = 0.00053). (**c**) Example single-trial activity of M2 and M1 communication subspace activity before and during reach initiation. (Left) Early learning. (Right) Late learning. (**d**) ROC analysis of detection of reach initiation from M2 and M1 communication subspace activity using logistic regression (example animal). (Inset) Difference in reach detection with learning quantified as the area under the curve (AUC) for all animals. (mixed effect model, r = 4 rats. 0.66 ± 0.03 for Early; 0.87 ± 0.02 for Late, $p = 6.63 \times 10^{-5}$). (**e**) Example single-trial prediction of reach initiation using the model built in (d). (Left) Early learning. (Right) Late learning. (**f**) Comparison of mean prediction of reach initiation during early (grey) and late (gold) learning as in (e). Mean of all trials for example animal. Quantified as difference between single-trial mean of pre-reach window and mean of reach start window signal. (mixed effect model; 0.02 ± 0.04 for Early; 0.33 ± 0.01 for Late, p < 0.0001).

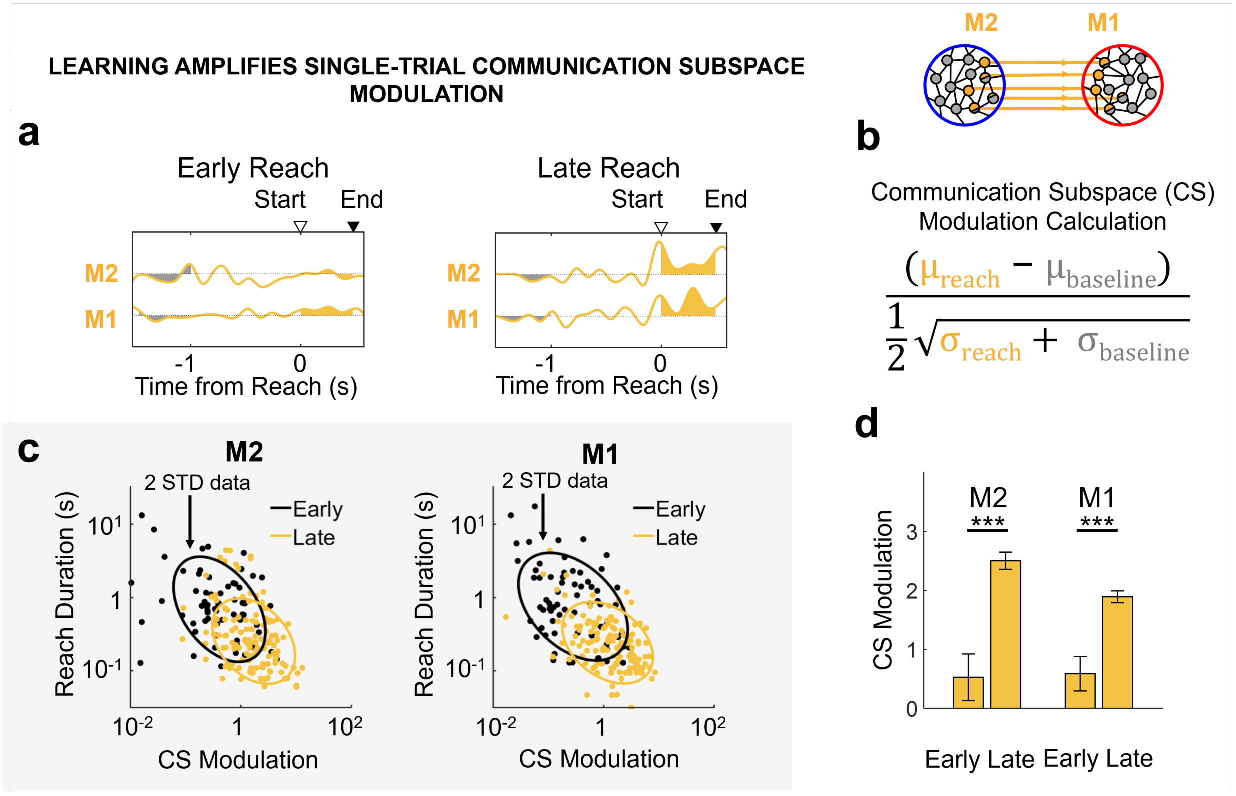


Figure. 1.5. Learning drives communication subspace encoding of reach duration.

(**a**) Example single trial M2 and M1 communication subspace activity in (left) early and (right) late learning. Reach duration is indicated by triangles marking reach start (open triangle) and reach end (filled triangle). (**b**) Equation for calculating reach modulation (see Methods). (**c**) Neural reach modulation predicts reach duration. Single-trial neural reach modulation for M2 (left) and M1 (right) communication subspace activity is plotted against single trial reach duration. Points show randomly subselected trials, with ellipses fitted to 2 standard deviations of the full dataset. All trials were used for quantification. Single-trial neural reach modulation and reach duration are significantly linearly related (mixed-effect model, M2: log slope = -0.27, p = 2.36 x 10^{-44}, M1: log slope = -0.23, p = 2.05 x 10^{-41}). (**d**) Reach modulation increases in both M1 and M2 communication subspace activity with learning (mixed effect model; M2: 0.53 ± 0.40 for Early; 2.60 ± 0.15 for Late, p = 2.48 x 10^{-43}, M1: 0.59 ± 0.29 for Early; 2.01 ± 0.10 for Late, p = 1.10 x 10^{-42}).

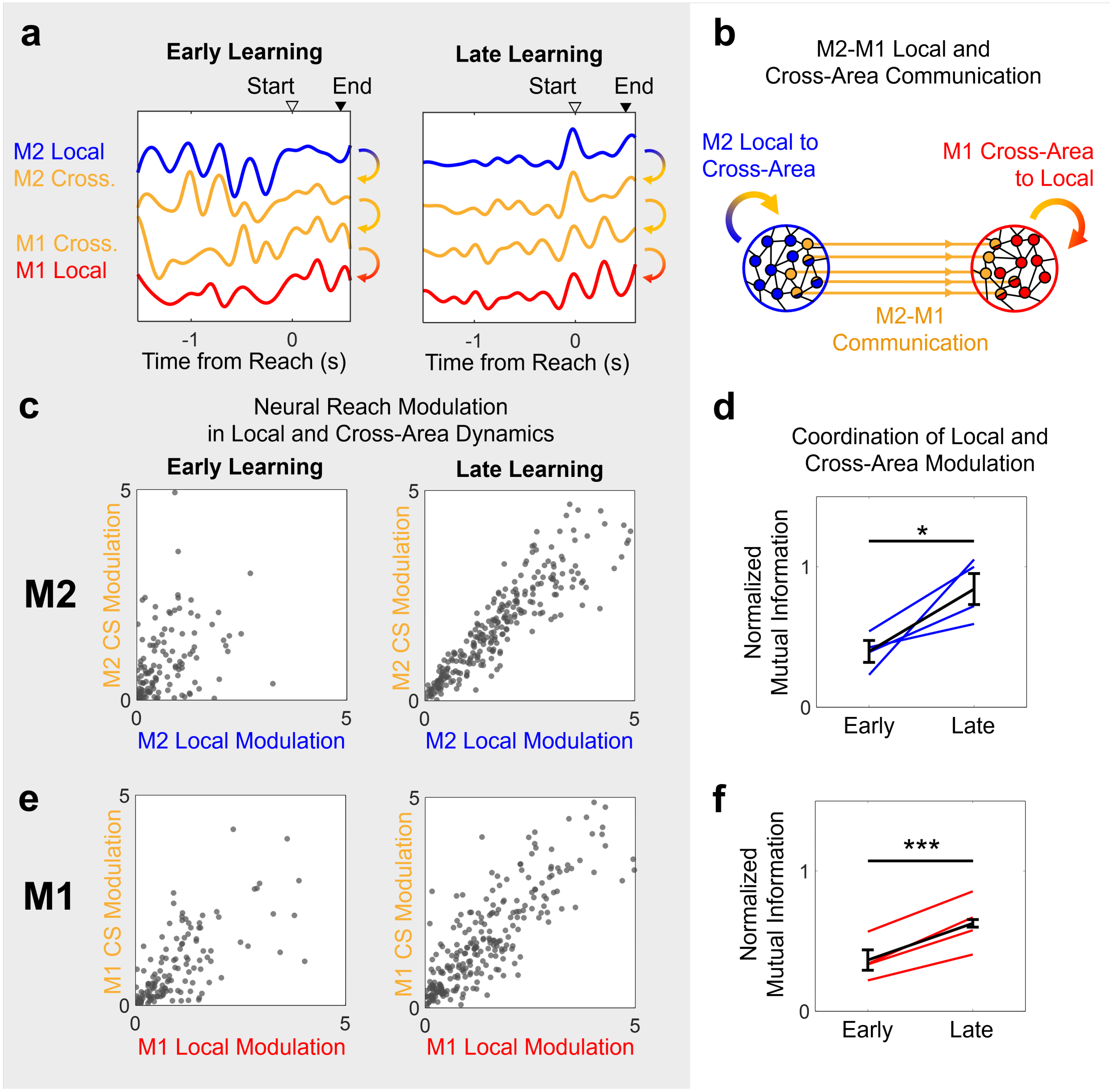

a
Early Learning
Late Learning
Start
End
M2 Local
M2 Cross.
M1 Cross.
M1 Local
-1
0
Time from Reach (s)
b
M2-M1 Local and
Cross-Area Communication
M2 Local to
Cross-Area
M1 Cross-Area
to Local
M2-M1
Communication
c
Neural Reach Modulation
in Local and Cross-Area Dynamics
Early Learning
Late Learning
M2
M2 CS Modulation
M2 Local Modulation
0
5
d
Coordination of Local and
Cross-Area Modulation
Normalized
Mutual Information
*
0
1
Early
Late
e
M1
M1 CS Modulation
M1 Local Modulation
0
5
f
Normalized
Mutual Information
***
0
1
Early
Late

Figure. 1.6. Learning increases information sharing between local and cross-area dynamics.

(**a**) Example single-trial activity from one animal's M2 local subspace, M2 and M1 communication subspaces, and M1 local subspace from (Left) early and (Right) late learning. (**b**) Model diagram of information flow from M2 local dynamics to M2-M1 cross-area dynamics, to M1 local dynamics. (**c**) Neural reach modulation (d') of single-trial neural activity in local versus communication subspace in M2 during (Left) early and (Right) late reach learning. Neural trajectories were first normalized to their maximum values as in (a) before calculating neural reach modulation. (**d**) Quantification of (**c**). Mutual information between single-trial modulation of local and cross-area dynamics increases with learning in M2 (mixed effect model; M2: 0.42 ± 0.08 for Early; 0.84 ± 0.11 for Late, p = 0.01). (**e**) As in (c) but for M1. (**f**) As in (d) but for M1. Mutual information between single-trial modulation of local and cross-area dynamics increases with learning in M1 (mixed effect model; M1: 0.39 ± 0.07 for Early; 0.64 ± 0.02 for Late, $p = 3.22 \times 10^{-5}$).

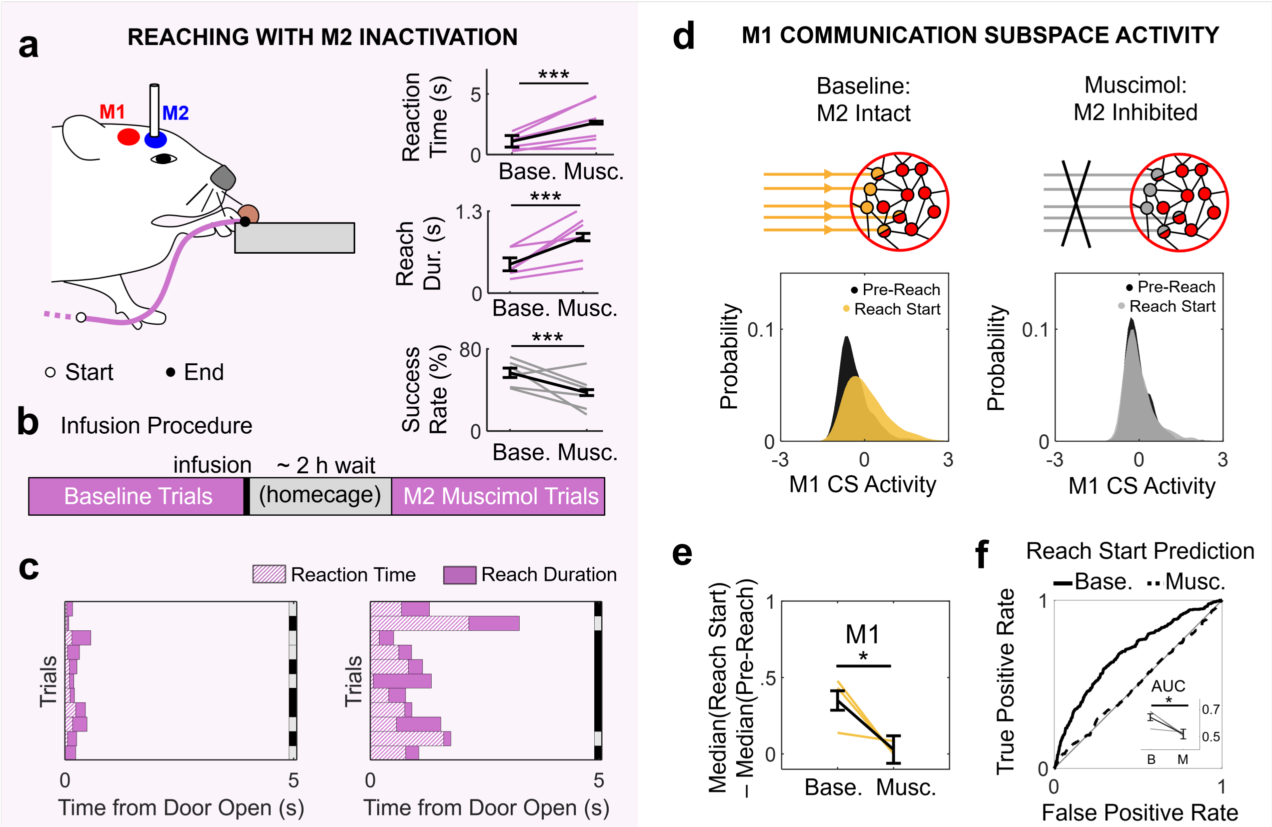


Figure. 1.7. M2 inhibition disrupts learned reach behavior and encoding of reach initiation in M1 cross-area activity.

(**a**) (Top left) Rats previously trained on the reach-to-grasp task were infused with muscimol in M2. (Top right) M2 inactivation increased reaction time (mixed effect model, 1.09 ± 0.48 for Baseline; 2.63 ± 0.12 for M2 Muscimol, p = 4.17 x 10^{-34}), increased reach duration (mixed effect model, 0.46 ± 0.10 for Baseline; 0.89 ± 0.06 for M2 Muscimol, p = 3.02 x 10^{-13}), and decreased success rate (mixed effect model, 56.78% ± 4.60 for Baseline; 37.43% ± 2.89 for M2 Muscimol, p = 3.62 x 10^{-11}). (**b**) Experimental paradigm for evaluation of reach behavior during M2 inactivation. (see Materials and Methods). (**c**) Example consecutive single-trial representations of reaction time and reach duration for baseline (left) and muscimol inactivation (right). Right border of plot shows accuracy, with success in gray and failure in black. (**d**) (Left) Neural activity from M1 communication subspace before (black) and during (yellow) reach initiation during baseline trials. M1 communication subspace neural weights were defined during baseline period and used to calculate neural activity during both baseline and M2 inactivation trials. (Right) As in Left, but during M2 inactivation trials. Activity during reach initiation is shown in grey. (**e**) Quantification of (d) as the difference between median pre-reach and reach activity during baseline and M2 inactivation trials in M1 communication subspace (mixed effect model, r = 3 rats. 0.35 ± 0.06 for Baseline; 0.03 ± 0.09 for M2 Muscimol, p = 0.02). (**f**) Detection of reach initiation from M1 communication subspace activity using ROC analysis (example animal). (Inset) Difference in reach detection quantified as the area under the curve (AOC) for all animals. (mixed effect model, r = 3 rats. 0.64 ± 0.03 for Baseline; 0.52 ± 0.04 for M2 Muscimol, p = 0.02).

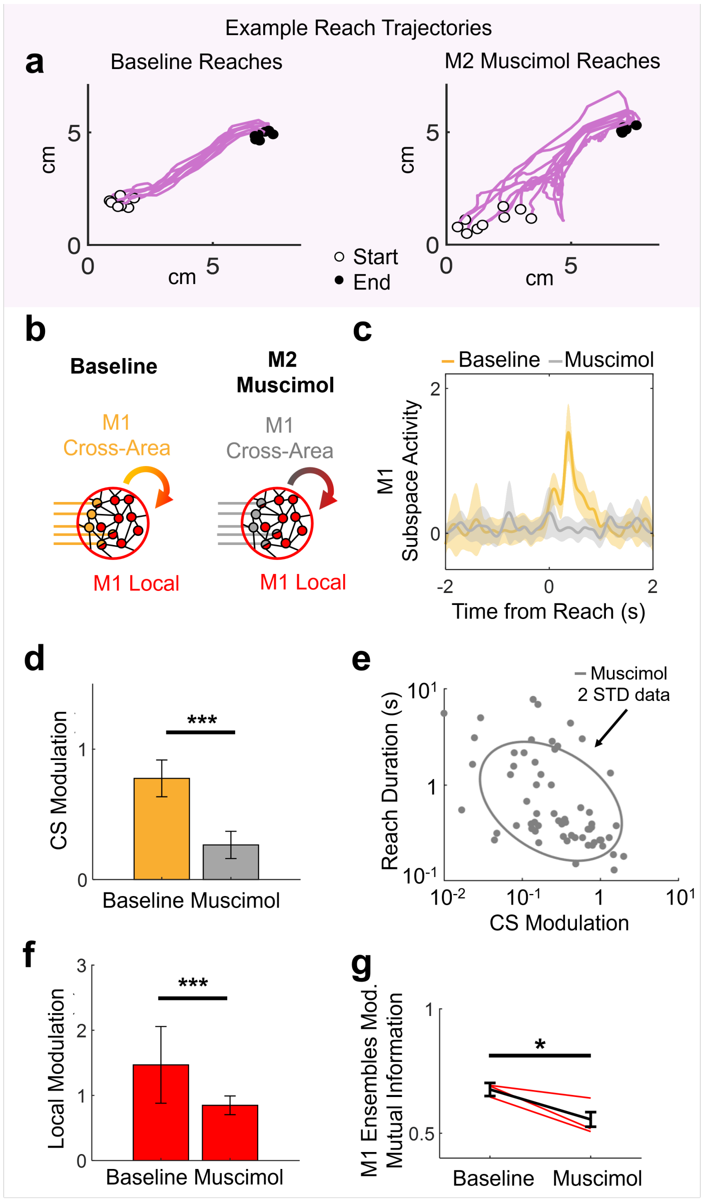

Example Reach Trajectories
a
Baseline Reaches
M2 Muscimol Reaches
cm
5
0
0
cm
5
○ Start
● End
b
Baseline
M2 Muscimol
M1 Cross-Area
M1 Cross-Area
M1 Local
M1 Local
c
Baseline
Muscimol
M1 Subspace Activity
2
0
-2
0
2
Time from Reach (s)
d
CS Modulation
***
1
0
Baseline Muscimol
e
Muscimol 2 STD data
Reach Duration (s)
10^1
1
10^{-1}
10^{-2}
10^{-1}
1
10^1
CS Modulation
f
Local Modulation
***
3
2
1
0
Baseline Muscimol
g
M1 Ensembles Mod. Mutual Information
*
1
0.5
Baseline
Muscimol

Figure. 1.8. M2 inhibition disrupts M1 encoding of reach duration.
(**a**) Example reach trajectories during (Left) Baseline trials and (Right) M2 Muscimol Inhibition trials. (**b**) Explanatory diagram showing hypothesis that M2 inactivation disrupts activity transmission between M1 cross-area and M1 local dynamics. (**c**) Mean M1 cross-area neural dynamics during baseline (yellow) M2 inactivation (grey) trials. M1 communication subspace neural weights were defined during baseline period and used to calculate communication subspace activity during both baseline and M2 inactivation trials. Shaded areas are 2 x standard error across trials. (**d**) M1 Communication Subspace (CS) neural modulation decreases significantly with M2 inactivation (mixed effect model, 0.78 ± 0.14 for Baseline; 0.27 ± 0.10 for M2 Muscimol, $p = 1.31 \times 10^{-6}$). (**e**) Single-trial M1 CS modulation predicts single-trial reach duration even during M2 inactivation (mixed effect model, log slope = -0.26, $p = 9.99 \times 10^{-8}$). Plot shows random subsampling of trials across animals, all trials were used in quantification. (**f**) M1 Local Subspace modulation decreases significantly with M2 inactivation (mixed effect model, 1.47 ± 0.59 for Baseline; 0.85 ± 0.14 for M2 Muscimol, $p = 1.64 \times 10^{-5}$). (**g**) Mutual Information between M1 local and communication subspace modulation decreases with M2 inactivation (mixed effect model, 0.67 ± 0.03 for Baseline; 0.56 ± 0.03 for M2 Muscimol, p = 0.01).

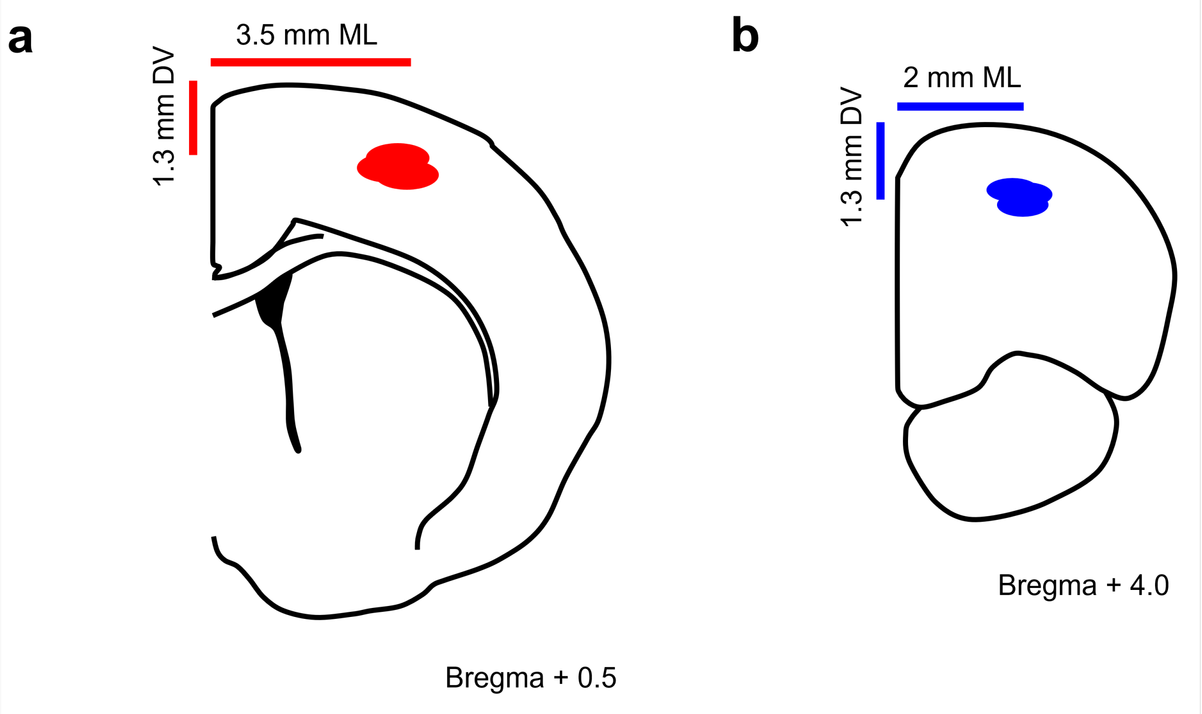


Figure. 1.S1. M1 and M2 electrode localization.

(**a**) Electrolytic lesion sites marking M1 electrode locations for three learning animals. (**b**) As in (a), but for M2.

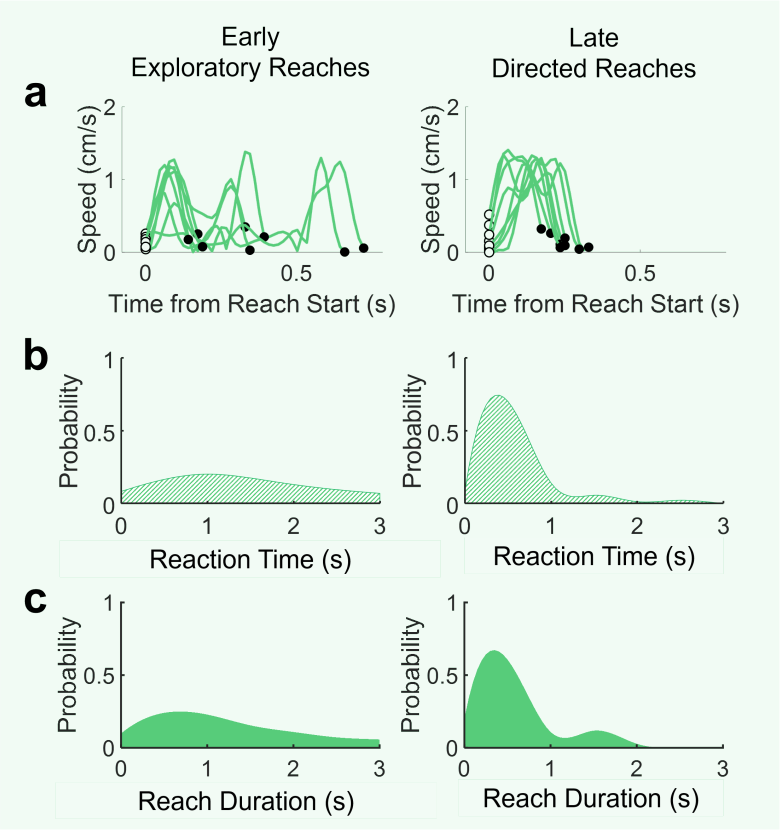


Figure. 1.S2. Elaboration of reach-to-grasp learning behavior.

(**a**) Speed profile for example trials in (Left) early exploratory reaches and (Right) late directed reached. Single-trial reach duration is driven by efficiency of reach targeting rather than maximal reaching speed. (**b**) Probability distribution of reaction times in (Left) early exploratory reaches and (Right) late directed reaches for all animals. (**c**) Probability distribution of reach durations in (Left) early exploratory reaches and (Right) late directed reaches for all animals.

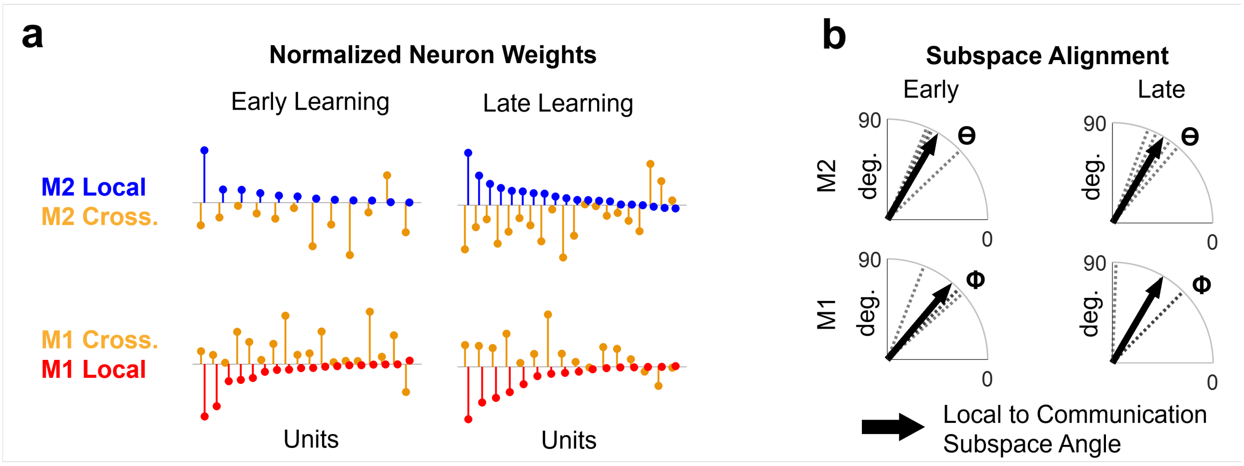


Figure. 1.S3. Motor behavior engages separable local and cross-area dynamics.

(**a**) Length of stems indicate weights for each neuron's contribution to local or cross-area activity, derived using PCA and CCA respectively. Neuron weights were normalized by the maximum value for any neuron in that subspace. PCA and CCA weights are shown offset and opposing for visual clarity; for all subspaces, most neurons had positive weights. M2 and M1 subspace neuron weights in (Left) early learning and (Right) late learning. Mutual information between PCA and CCA weights did not change with learning (mixed effect model, M2: 0.89 ± 0.07 for Early; 0.78 ± 0.10 for Late, p = 0.29, M1: 0.79 ± 0.05 for Early; 0.77 ± 0.07 for Late, p = 0.88). (**b**) Angle in multi-dimensional space between local and communication subspaces. Black arrow is the mean angle across animals (n = 4), dashed lines show values for each animal. In order, M2 (**Θ**) and M1 (**Φ**) subspaces angles in (Left) early learning and (Right) late learning. In each region, the angles between local and cross-area activity axes are significantly different from zero (mixed effect model, M2: p = 7.92 x 10^{-5}, M1: p = 6.75 x 10^{-5}), are not significantly different between early and late learning (mixed effect model, M2: 66.15 deg ± 6.99 for Early; 69.10 ± 6.26 for Late, p = 0.23, M1: 63.65 ± 6.54 for Early; 75.72 ± 9.25 for Late, p = 0.65).

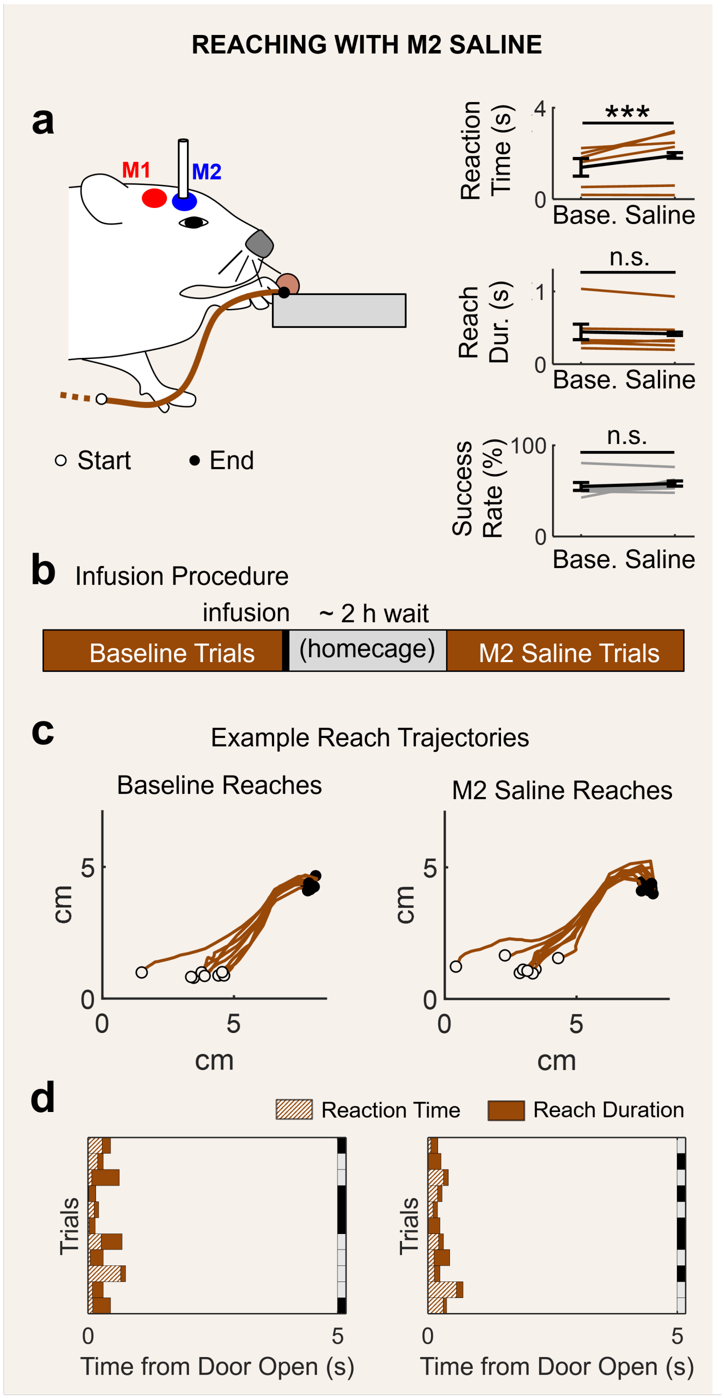

REACHING WITH M2 SALINE
a
M1
M2
○ Start
● End
Reaction Time (s)
4
0
***
Base. Saline
Reach Dur. (s)
1
0
n.s.
Base. Saline
Success Rate (%)
100
0
n.s.
Base. Saline
b
Infusion Procedure
infusion
~ 2 h wait
Baseline Trials
(homecage)
M2 Saline Trials
c
Example Reach Trajectories
Baseline Reaches
M2 Saline Reaches
cm
5
0
0
5
cm
d
▨ Reaction Time
■ Reach Duration
Trials
0
5
Time from Door Open (s)

Figure. 1.S4. M2 saline infusions do not affect learned reach behavior.

(**a**) (Top left) Rats previously trained on the reach-to-grasp task were infused with saline in M2. (Top right) M2 saline did not change reaction time (mixed effect model, 1.40s ± 0.38 for Baseline; 1.91s ± 1.12 for M2 Saline, $p = 2.10 \times 10^{-5}$), reach duration (mixed effect model, 0.44s ± 0.11 for Baseline; 0.42s ± 0.02 for M2 Saline, $p = 0.25$), or success rate (mixed effect model, 54.95% ± 4.32 for Baseline; 58.13% ± 2.83 for M2 Saline, $p = 0.26$). (**b**) Experimental paradigm for evaluation of reach behavior during M2 saline infusion. (**c**) Example reach from a single animal during (Left) baseline and (Right) M2 saline infusion. (**c**) Example consecutive single-trial representations of reaction time and reach duration. Right border of plot shows accuracy, with success in gray and failure in black.

References

1. Sporns, O., Chialvo, D., Kaiser, M. & Hilgetag, C. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences* **8**, 418–425 (2004).

2. Cao, V. Y. *et al.* Motor Learning Consolidates Arc-Expressing Neuronal Ensembles in Secondary Motor Cortex. *Neuron* (2015). doi:10.1016/j.neuron.2015.05.022

3. Hikosaka, O., Nakamura, K., Sakai, K. & Nakahara, H. Central mechanisms of motor skill learning. *Current Opinion in Neurobiology* **12**, 217–222 (2002).

4. Makino, H. *et al.* Transformation of Cortex-wide Emergent Properties during Motor Learning. *Neuron* **94**, 880-890.e8 (2017).

5. Perich, M. G., Gallego, J. A. & Miller, L. E. A Neural Population Mechanism for Rapid Learning. *Neuron* **100**, 964-976.e7 (2018).

6. Tanji, J. Sequential Organization of Multiple Movements: Involvement of Cortical Motor Areas. *Annual Review of Neuroscience* **24**, 631–651 (2001).

7. Athalye, V. R., Ganguly, K., Costa, R. M. & Carmena, J. M. Emergence of Coordinated Neural Dynamics Underlies Neuroprosthetic Learning and Skillful Control. *Neuron* **93**, 955-970.e5 (2017).

8. Ganguly, K., Dimitrov, D. F., Wallis, J. D. & Carmena, J. M. Reversible large-scale modification of cortical networks during neuroprosthetic control. *Nature Neuroscience* **14**, 662–667 (2011).

9. Kawai, R. *et al.* Motor Cortex Is Required for Learning but Not for Executing a Motor Skill. *Neuron* **86**, 800–812 (2015).

10. Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).

11. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

12. Arce-McShane, F. I., Ross, C. F., Takahashi, K., Sessle, B. J. & Hatsopoulos, N. G. Primary motor and sensory cortical areas communicate via spatiotemporally coordinated networks at multiple frequencies. *Proceedings of the National Academy of Sciences* **113**, 5083–5088 (2016).

13. Benchenane, K. *et al.* Coherent Theta Oscillations and Reorganization of Spike Timing in the Hippocampal- Prefrontal Network upon Learning. *Neuron* **66**, 921–936 (2010).

14. DeCoteau, W. E. *et al.* Learning-related coordination of striatal and hippocampal theta rhythms during acquisition of a procedural maze task. *Proceedings of the National Academy of Sciences* **104**, 5644–5649 (2007).

15. Koralek, A. C., Jin, X., Long Ii, J. D., Costa, R. M. & Carmena, J. M. Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills. *Nature* **483**, 331–335 (2012).

16. Loonis, R. F., Brincat, S. L., Antzoulatos, E. G. & Miller, E. K. A Meta-Analysis Suggests Different Neural Correlates for Implicit and Explicit Learning. *Neuron* **96**, 521-534.e7 (2017).

17. Chen, T.-W., Li, N., Daie, K. & Svoboda, K. A Map of Anticipatory Activity in Mouse Motor Cortex. *Neuron* **94**, 866-879.e4 (2017).

18. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience* **17**, 440 (2014).

19. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).

20. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. 13 (2019).

21. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat Neurosci* **17**, 1500–1509 (2014).

22. Pang, R., Lansdell, B. J. & Fairhall, A. L. Dimensionality reduction in neuroscience. *Current Biology* **26**, R656–R660 (2016).

23. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical Areas Interact through a Communication Subspace. *Neuron* (2019). doi:10.1016/j.neuron.2019.01.026

24. Gulati, T., Ramanathan, D. S., Wong, C. C. & Ganguly, K. Reactivation of emergent task-related ensembles during slow-wave sleep after neuroprosthetic learning. *Nat Neurosci* **17**, 1107–1113 (2014).

25. Lara, A. H., Cunningham, J. P. & Churchland, M. M. Different population dynamics in the supplementary motor area and motor cortex during reaching. *Nature Communications* **9**, 2754 (2018).

26. Laubach, M., Wessberg, J. & Nicolelis, M. A. L. Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. *Nature* **405**, 567–571 (2000).

27. Laurent, G. Olfactory network dynamics and the coding of multidimensional signals. *Nature Reviews Neuroscience* **3**, 884–895 (2002).

28. Lin, I.-C., Okun, M., Carandini, M. & Harris, K. D. The Nature of Shared Cortical Variability. *Neuron* **87**, 644–656 (2015).

29. Mackevicius, E. L. *et al.* Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *eLife* **8**, e38471 (2019).

30. Ramanathan, D. S., Gulati, T. & Ganguly, K. Sleep-Dependent Reactivation of Ensembles in Motor Cortex Promotes Skill Consolidation. *PLoS Biol* **13**, e1002263 (2015).

31. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience* **36**, 337–359 (2013).

32. Yttri, E. A. & Dudman, J. T. Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature* **533**, 402–406 (2016).

33. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience* **17**, 1784–1792 (2014).

34. Whishaw, I. Q. & Pellis, S. M. The structure of skilled forelimb reaching in the rat: A proximally driven movement with a single distal rotatory component. *Behavioural Brain Research* **41**, 49–59 (1990).

35. Wong, C. C., Ramanathan, D. S., Gulati, T., Won, S. J. & Ganguly, K. An automated behavioral box to assess forelimb function in rats. *Journal of Neuroscience Methods* **246**, 30–37 (2015).

36. Whishaw, I. Q., Pellis, S. M., Gorny, B. P. & Pellis, V. C. The impairments in reaching and the movements of compensation in rats with motor cortex lesions: an endpoint, videorecording, and movement notation analysis. *Behavioural Brain Research* **42**, 77–91 (1991).

37. Darling, W. G., Pizzimenti, M. A. & Morecraft, R. J. Functional Recovery Following Motor Cortex Lesions in Non-Human Primates: Experimental Implications for Human Stroke Patients. *J Integr Neurosci* **10**, 353–384 (2011).

38. Rothwell, P. E. *et al.* Input- and Output-Specific Regulation of Serial Order Performance by Corticostriatal Circuits. *Neuron* **88**, 345–356 (2015).

39. Fries, P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences* **9**, 474–480 (2005).

40. Svoboda, K. & Li, N. Neural mechanisms of movement planning: motor cortex and beyond. *Current Opinion in Neurobiology* **49**, 33–41 (2018).

41. Golub, M. D. *et al.* Learning by neural reassociation. *Nature Neuroscience* (2018). doi:10.1038/s41593-018-0095-3

42. Dechery, J. B. & MacLean, J. N. Emergent cortical circuit dynamics contain dense, interwoven ensembles of spike sequences. *Journal of Neurophysiology* **118**, 1914–1925 (2017).

43. Jazayeri, M. & Afraz, A. Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017).

44. Otchy, T. M. *et al.* Acute off-target effects of neural circuit manipulations. *Nature* **advance online publication**, (2015).

45. Shima, K. & Tanji, J. Neuronal Activity in the Supplementary and Presupplementary Motor Areas for Temporal Organization of Multiple Movements. *Journal of Neurophysiology* **84**, 2148–2160 (2000).

46. Saiki, A. *et al.* Different Modulation of Common Motor Information in Rat Primary and Secondary Motor Cortices. *PLoS ONE* **9**, e98662 (2014).

47. Smith, N. J., Horst, N. K., Liu, B., Caetano, M. S. & Laubach, M. Reversible inactivation of rat premotor cortex impairs temporal preparation, but not inhibitory control, during simple reaction-time performance. *Front. Integr. Neurosci.* **4**, 124 (2010).

48. Quiroga, R. Q., Nadasdy, Z. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput* **16**, 1661–1687 (2004).

49. Chung, J. E. *et al.* A Fully Automated Approach to Spike Sorting. *Neuron* **95**, 1381-1394.e6 (2017).

50. Ghahramani, Z. & Hinton, Geoffrey. The EM Algorithm for Mixtures of Factor Analyzers. 8 (1997).

51. Toutenburg, H. Everitt, B. S.: Introduction to Latent Variable Models. Chapman and Hall, London 1984. 107 pp., £ 9.50. *Biometrical Journal* **27**, 706–706 (1985).

52. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28**, 321–377 (1936).

53. Macmillan, N. A. & Creelman, C. D. *Detection theory: a user's guide*. (Lawrence Erlbaum Associates, 2005).

Chapter 2: Cross-Area Cortical Engagement During Brain-Machine Interface Learning

Authors: T. L. Veuthey[1,2,3,4]†, K. Derosier[1,3,4]†, K. Ganguly[3,4].

Affiliations:

[1]Neuroscience Graduate Program, University of California San Francisco, San Francisco CA.

[2]Medical Scientist Training Program, University of California San Francisco, San Francisco CA.

[3]Neurology and Rehabilitation Service, San Francisco Veterans Affairs Medical Center, San Francisco CA, USA.

[4]Department of Neurology, University of California San Francisco, San Francisco CA, USA.

† These authors contributed equally to this work.

Article Type: Original Research

Keywords: Motor Cortex, Premotor Cortex, Electrophysiology, Brain-Machine Interface

Abstract

Skilled movements are the foundation of our ability to interact with the environment. Learning skilled movements requires multiple cortical regions, such as the primary motor (M1) and premotor cortices (M2). While M2 is hypothesized to provide top-down guidance to M1, the functional relationship between these regions is still unclear. M1 and M2 are often studied in well-trained animals performing motor skills, but in this context, it is difficult to discern whether M2 activity is shaping M1 activity or sending signals directly to muscles. BMIs offer a way to disambiguate these alternatives. By using just M1 to control the BMI, we can effectively select M1 as the sole output of the cortical motor system. Unlike natural motor learning, where M2 may be necessary because of its direct motor output, in M1 neuroprosthetic learning, M2 can only be necessary via its influence on M1. Using a M1-BMI learning task during which we simultaneously recorded in M1 and M2, we show that M2 neurons are modulated during M1-BMI learning. Additionally, the proportions of M2 neurons modulated are not significantly different that of M1-indirect neurons, which also do not contribute to direct BMI control. These results suggest that BMI learning engages plasticity similarly in top-down regions in the functional motor network as in local neural population.

Introduction

The functional motor network is made of many interconnected cortical and sub-cortical regions[1]. However, the complexity of these connections makes it difficult to assign causality between specific neural activity and natural behaviors. Brain-machine interfaces (BMIs) provide a tractable approach to understanding the role of populations of neurons in behavior. Using BMIs, we can *create* a causal link between specific patterns of population neural activity and the behavior of an external actuator [2]. Many BMI studies have focused on understanding how different aspects of neural activity within single brain regions affect learning[3–10], specifically probing the role of intrinsic connectivity patterns[9–11] (i.e. 'manifolds'), sleep[7,12], and cognitive strategies in BMI learning[8]. While understanding parameters of neural activity within single brain regions is critical, this approach cannot tell us how activity coordinated across different brain regions contributes to learning and behavior. To address this, it is necessary to analyze simultaneous activity from many interacting regions during learning[3,13].

In the motor system, cross-area interactions are hypothesized to play an important role in learning[1,14]. However, studies about interactions between regions during learning are rare[15–17]. Findings from these studies are also confounded by the known redundancies in motor network projection patterns; for example, forelimb regions in premotor (M2) and motor cortex (M1) both send projections to the same segments in the spinal cords[18]. Therefore, it becomes difficult to dissociate the roles of M1 and M2 activity during forelimb movements[19,20]. The prominent anatomic connections between M1 and M2 [18,21] also make lesion and inactivation studies difficult to interpret, as chronic and acute interventions can lead to disruptions beyond the inhibited region[22,23] as well as motor map changes and functional compensation within intact regions[24,25]. Some studies of M1 and M2 during motor skill learning point to top-down guidance of M1 by M2[15,21], but other studies suggest that the two areas function in parallel[26,27]. Brain-machine interfaces (BMIs) have the potential to clarify this relationship by simplifying the link between neural activity and behavioral output (Fig. 2.1). By using a BMI task, we can probe the connection between M1 and M2 during M1-driven BMI learning. We hypothesize that, although M2 neurons do not contribute directly to M1-BMI performance, M2 provides top-down guidance for M1-BMI learning.

This hypothesis predicts several specific findings. First, M2 neurons must be task-modulated during M1-BMI learning. Without task-specific activity, it is difficult to imagine how M2 neurons would contribute to M1-BMI control. Second, M2 neural activity must predict M1 activity in order to have a causal influence. Third, disruption of M2 during M1-BMI learning must impair learning. Here we provide preliminary evidence for the first prediction, and outline an approach for future analyses and M2 manipulations during M1-BMI learning.

Results

Four rats were implanted with grids of microelectrodes in M2 and M1 and then trained to use M1 neural signals to directly control the angular velocity of a mechanical actuator that delivered water rewards [7] (Fig. 2.2a). A linear decoder converted the firing rates of two sets of neurons (hereafter referred to as direct unit pools) into the angular velocity of the actuator (see Materials and Methods). In three of the four rats, the decoder additionally provided visual feedback in the form of a grey circular disk whose location on a diagonal line indicated neural state (Fig. 2.2a). We also recorded activity of M1 and M2 neurons whose activity was not linked to actuator movements (hereafter referred to as M1 indirect units and M2 units, respectively). The decoder was fixed during each daily session; consequently, improvements in task performance were exclusively due to neural learning mechanisms.

Each trial started with an auditory cue coinciding with the opening of a door allowing access to the water spout. If rats achieved the neural firing rate target, success was indicated with an auditory cue, and water reward was delivered via a metal spout through the slot. If rats failed to achieve the target in the set time, failure was indicated with an auditory cue and the closing of the gate, followed by a timeout period. Task performance was monitored through two complementary metrics: success rate and trial duration. While success rate is a binary measure of the rats' ability to achieve the target neural activity sometime during each attempt, trial duration provides a continuous measure of rats' ability to modulate their neural activity quickly on reaction to trial cueing. Over the course of a typical 2-h robust learning session, rats' performance improved in accordance with both measures (Fig. 2.2b).

M2 neurons are engaged by M1-BMI learning

Preliminary neural data (n = 2) confirms past studies[6,7,12] showing that as rats learn the task, both direct and indirect M1 neurons are modulated (Fig. 2.3a). Importantly, many M2 units also become task-modulated (Fig. 2.3b), suggesting that, even though no muscle movement is required, intrinsic M2-M1 communication drives M2 cortical engagement during M1-BMI learning. We quantified each neuron's task modulation by comparing deviations in the peri-event time histogram (PETH) of each neuron to those of 10,000 artificially-created PETHs in which activity of each trial was circularly shuffled to eliminate task-based activity alignment[28] (see Materials and Methods). Using this approach, we found that 44% of M1-indirect neurons (28 of 63) and 54% of M2 neurons (35 of 65) were modulated at the end of trials, which is the timepoint at which neural activity determined trial success. These proportions are not significantly different (X^2 test, t(1) = 1.13, p = 0.28), suggesting that BMI learning engages plasticity similarly in both local and distant populations of neurons.

Discussion

This study uses brain-machine interfaces as a novel approach to understanding interactions between two regions in a functional neural system. We analyzed engagement of M2 neurons during an M1-driven BMI learning task, and compared M2 modulation to that of M1 indirect neurons, neither of which are required for BMI control. Surprisingly, we found that similar proportions of M2 neurons and M1 indirect neurons are engaged by M1-BMI learning, suggesting that task-based cross-area M2-M1 communication and M1-M1 local communication are comparable, despite much higher local anatomic connectivity.

BMI as a tool for circuit dissection

Although brain machine interfaces are often thought of in terms of therapeutic neuroprosthetics, there is also a long history of using BMIs to better understand neural circuits and learning. The ability of the brain to adapt and learn to control a fixed decoder BMI has been shown in multiple brain regions and in both rodents and primates [2,4,6,12,13,29–31]. As with natural movement, proficient BMI control is associated with stereotyped patterns of activity[4,6]. BMI learning is also sleep-dependent [12] and requires corticostriatal plasticity [13]. These concordances between natural movement and BMI learning suggest that BMI can be used as a tool to understand principles and mechanisms of learning which are not unique to the BMI control. However, BMI decoder design allows experimenters to determine which neurons are directly causal, opening up new avenues for analysis of mechanisms of plasticity which drive task performance.

Do M2 neurons directly contribute to M1-BMI control?

Prior work has shown that, while modulation of M1-direct units increase with learning, modulation of M1-indirect neurons decreases with learning[6] in a sleep-dependent manner [12]. As of yet, it is unclear whether M2 neuron modulation increases or decreases with multi-session learning and sleep. This dissociation may allow us to better understand whether M2 neurons directly contribute to M1-BMI control. An increase in modulation (similar to M1-direct neurons) would suggest a causal influence on M1-direct neurons, while a decrease in modulation (similar to M1-indirect neurons) would suggest that initial M2 neural modulation reflects exploratory strategies during early learning which are then culled as learning progresses.

Do M2 neurons contribute to M1-BMI learning?

While M2 is required for skilled motor learning [32], in natural learning it is impossible to dissociate M2's role in direct control of muscles from its role in top-down guidance of M1 and other regions in the motor network during learning. BMI experiments provide an approach for testing whether M2 activity during early exploration is necessary for M1-BMI learning. Using retrograde viral vectors containing inhibitory opsins [33], it is possible to infect cells from distant regions which specifically project axons to the injection site. Injecting such viruses in M1 would allow optogenetic access to M2 cells which project to M1. Subsequently inhibiting either the M2 cell bodies or synaptic terminals allows for specific manipulation of M2 to M1 signals. As BMI tasks require experimenters to choose the neurons which drive the task, it would be possible to select M1-direct neurons whose firing rates are affected by manipulation of M2 neural activity. By manipulating M2 neural activity during M1-BMI learning, we would then be able to interpret changes in M1-BMI learning as stemming from disrupted M2 inputs. This approach is only viable because, unlike many natural motor tasks, rats are able to learn a new BMI decoder within a single day [7], permitting both same-day tracking of neurons throughout the learning process and the potential for many days of repeated learning experiments using the same task.

How are M2 and M1 population interactions coordinated?

Chapter 1 used computational analysis methods to understand the coordination of local and cross-area M1-M2 population dynamics during learning of a skilled motor behavior, the reach-to-grasp task. We found that local and cross-area population representations of skilled reaching became more similar with learning, suggesting that cross-area dynamics participate in coordination and transformation of activity between nodes of the motor network. We propose that this transformation of activity takes place in a distributed manner across many neurons in both M1 and M2. However, we did not touch on how the signals in both regions become temporally synced. To do this, we must employ analytic methods which can take into account the temporal relationships in population activity, rather than assuming that streams of binned activity are independent samples. Studies in non-human primates and rodents indicate that M2 is particularly important for sequence learning but not learned sequence execution [34], suggesting a dynamic relationship between M2 and M1 during sequence learning. One hypothesis is that M2 temporally binds and organizes [15,35] motor network neural activity, and that M1 activity represents motor primitives whose expression drives submovements [36,37]. In contrast, studies in songbirds have found that multiple adult neural sequences corresponding to song syllables emerge from the growth and splitting of a common precursor neural sequence[38]. Consequently, both binding and splitting of neural sequences seem like viable mechanisms for developing new motor skills. By tracking the relative timing and stability of neural sequences in M1 and M2 during M1-BMI learning, future experiments may provide a less effector-dependent view on how learning drives robust temporal relationships in neural activity (i.e. the 'tiling’ of activity) to produce skilled behaviors. Being able to track the functional interactions of causally defined M1-direct neurons with both M1-indirect and M2 neurons may also suggest a parameters space for inter-neuron activity coordination to use in biologically plausible in silico modeling of interacting neural networks.

Conclusions

Many functions of M2 have been proposed, including preparation of upcoming actions[39], orchestration of sequential movements[1], and new learning of skilled actions [32]. These different functions can be united in the context of the dynamical systems model of motor cortex. There is a growing body of work suggesting that M1 can be understood as a dynamical system [29,30,40,41], meaning that it has an internal drive [30] governing how future neural states evolve from past neural states. In this view, the role of M2 input might be to push the neural state in a different direction, against the established internal dynamics of M1. Movement preparation, unfamiliar movements, and transitions between sequence elements might all require M2 input because they are times when producing the correct behavior requires pushing against ongoing default dynamics of

M1. BMI tasks provide a unique opportunity for training M1 to produce specific dynamics and analyze how interactions between M1 and M2 contribute to that process.

Materials and Methods

Animal Care

All procedures were in accordance with protocols approved by the Institutional Animal Care and Use Committee at the San Francisco Veterans Affairs Medical Center. Adult male Long Evans rats (n = 4, 250–400 g; Charles River Laboratories) were housed in a 12-h/12-h light–dark cycle. All experiments were done during the light cycles. Rats were housed in groups of 2 animals prior to surgery and individually after surgery.

Surgery

All surgical procedures were performed using a sterile technique under 2–4% isoflurane. Surgery involved cleaning and exposure of the skull, preparation of the skull surface (using cyanoacrylate) and then implantation of the skull screws for overall headstage stability. Reference screws were implanted posterior to lambda and ipsilateral to the neural recordings. For experiments involving physiological recordings, craniotomy and durectomy were performed, followed by implantation of the neural probes. For experiments involving only infusions, burr holes were drilled in the appropriate locations, followed by implantation of the cannulas. Postoperative recovery regimen included the administration of 0.02 mg per kg body weight buprenorphine for 2 days, and 0.2 mg per kg body weight meloxicam, 0.5 mg per kg body weight dexamethasone and 15 mg per kg body weight trimethoprim sulfadiazine for 5 days. All animals were allowed to recover for 1 week prior to further behavioral training.

Electrode array implants

Rats were implanted with two 32-channel tungsten wire probes (TDT or Innovative Neurophysiology), one each in M1 (+0.5 AP, +3.5 ML, -1.5 DV) and M2 (+4.0 AP, +1.5 ML, -1.5 DV), contralateral to reaching arm. Infusion cannulas were implanted in M2 (+4.0 AP, +1.5 ML, -1.5 DV) for infusion-only animals. For rats with both M2 electrode arrays and cannulas, the cannula was attached to the electrode array prior to surgery.

Histology

Final placement of the electrodes was monitored online based on implantation depth and verified histologically at the end of the experiments. Rats were anesthetized with isoflurane and transcardially perfused with 0.9% sodium chloride, followed by 4% formaldehyde. The harvested brains were post-fixed for 24 h and immersed in 20% sucrose for 2 days. Coronal cryostat sections (40-μm thickness) were mounted with permount solution (Fisher Scientific) on superfrosted coated slides (Fisher Scientific). Images of a whole section were taken by a HP scanner, and microscope images were taken by a Zeiss microscope.

Electrophysiology data collection

We recorded extracellular neural activity using tungsten microwire electrode arrays (MEAs, n = 7 rats, TDT or Innovative Neurophysiology). We recorded spike and LFP activity using a 128–channel TDT–RZ2 system (TDT). Spike data was sampled at 24,414 Hz and LFP data at 1,018 Hz. Analog headstages with a unity gain and high impedance (~1 GΩ) were used. Snippets of data that crossed a high signal-to noise threshold (4 standard deviations away from the mean) were time-stamped as events, and waveforms for each event were peak aligned. MEA recordings were sorted offline using a density-based clustering algorithm (Mountainsort[42]). Clusters interpreted to be noise were discarded, but multi-units were kept for analysis. Trial-related timestamps (i.e., trial onset, trial completion, removal of pellet from pellet holder, and timing of video frames) were sent to the RZ2 analog input channel using an Arduino digital board and synchronized to neural data.

General brain-machine interface paradigm

Rats were trained using an automated behavior box, with components controlled by Matlab R2015a and an Arduino running the Adafruit Motor Library V1. Within the box, rats were unrestrained. Neural data was recorded and sorted online using software from Tucker Davis Technologies: for spout BMI, the software used was OpenEx; for visual BMI, it was Synapse. Spike counts from online sorting were imported into Matlab and used to control the feedback stimuli (see “Spout BMI” and “Visual BMI” for details). Trials started with an auditory cue and the opening of the plastic gate covering a slot in the back of the behavior box. When rats achieved the neural firing rate target, success was indicated with an auditory cue, and water reward was delivered via a metal spout through the slot. If rats failed to achieve the target in the set time, failure was indicated with an auditory cue and the closing of the gate, followed by a timeout period. The maximum trial length and the timeout period following failures were both manipulated over the course of the experiments to encourage learning, and ranged from 10-20s and 5-10s respectively.

Spout BMI

The spout BMI paradigm was used to train n = 1 rat. In this paradigm, feedback about progress to the firing rate target was given via the movement of the water spout used for reward. Eight "direct" units were chosen based on having good signal-to-noise and neither unusually high nor unusually low firing rates. Of the direct units, 4 units were arbitrarily assigned to the "positive pool", and 4 units were arbitrarily assigned to the "negative pool". The same channels were used for all sessions, but we did not directly test for unit similarity across days. At the beginning of each session, a 30 minute baseline recording was taken and used to fit mean firing rates for each unit. During the task, for every 100ms bin, direct unit firing rates were computed, mean subtracted, and summed within pools. The "neural state" was computed as $s = g * (p - n)$, where p is the firing rate of the positive pool, n is the firing rate of the negative pool, and g is an experimenter-controlled gain parameter. The neural state was smoothed by averaging it with its previous value, and then used to control the position of the water spout, such that increasing the difference between p and n moved the spout towards the rat. Once the spout crossed a threshold value, the trial was considered a success.

Visual BMI

The visual BMI paradigm was used to train n = 3 rats. In this paradigm, feedback about progress to the firing rate target was given via the movement of both a visual cue on a computer monitor placed outside the behavior box and of the water spout used for reward. 4-8 "direct" units were chosen based on having good signal-to-noise and neither unusually high nor unusually low firing rates. Of the direct units, 2-4 units were arbitrarily assigned to the “positive pool”, and 2-4 units were arbitrarily assigned to the “negative pool”. The same channels were used for all sessions, but we did not directly test for unit similarity across days. At the beginning of each session, a 5-10 minute baseline recording was taken. The baseline data was divided into overlapping 100ms bins. For every bin, firing rates were summed within the positive and negative pools, and the difference between the two pools was computed. Gamma distributions were fitted to the histogram of firing rate differences using the Matlab function fitdist. During the task, for every 100ms bin, firing rates were summed within the positive and negative pools, and the difference between the two pools was computed. This difference was fed into the cumulative distribution function of the baseline distribution to obtain the “neural state”. When the neural state crossed an experimenter-defined threshold, the trial was considered a success. Typical threshold values were 0.85 - 0.95.

The neural state was smoothed by averaging it with its previous value, and then used to give rats feedback in two ways. First, a computer monitor outside the behavior box displayed a circular “cursor” that moved along a line towards a stationary “target” circle. The position of the cursor along the line was a direct readout of neural state, moving from the top left to the bottom right of the screen (i.e. closer to the rat) as neural state increased. Second, the neural state was also used to control the position of the water spout. The angular speed of the water spout was limited to 1 degree/s, but otherwise the position of the spout was proportional to the neural state such that as neural state increased, the spout moved closer to the rat.

Neural Analysis: Single-Unit Task Modulation

We will calculate task-relevant modulation using a bootstrap circular shuffle test[28]. Prior work in our lab on shows that M2 units may be positively, negatively, or multiphasically modulated during a natural reaching task, so it is important to use a quantification method that accounts for all of these possibilities. The circular shuffle test creates a set of surrogate peri-event time histograms (PETHs) by adding random time jitter to each trial in a spike raster matrix. If the true PETH lies outside of the distribution of shuffled histograms, then that unit is significantly modulated by the trial events. In our preliminary analyses, we binned spikes at 10ms, smoothed using a 70ms Gaussian kernel, and used 10000 bootstrap samples.

Acknowledgements

We thank C. Campillo Rodriguez for assistance with histology; S.J. Won for assistance with perfusions and imaging. Funding: This work was supported by fellowship awards from the National Defense Science and Engineering Graduate Fellowship (NDSEG, https://ndseg.asee.org/) the UCSF Discovery Fellows Program, and the Markowski-Leach Fellowship, the UCSF Medical Scientist Training Program (to T.L.V.), and the UCSF Neuroscience Graduate Program (to T.L.V. and K.D.). Additional funds come from the Department of Veterans Affairs, Veterans Health Administration (VA Merit: 1I01RX001640 to K.G) and National Institute of Mental Health, NIH (5R01MH111871 to K.G.); and start-up funds from the UCSF Department of Neurology to K.G.. K.G. also holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund (1009855) and an Independent Scientist Award (1K02NS093014) from the National Institute of Neurological Disorders and Stroke, NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Author contributions: T.L.V., K.D. and K.G. designed the study. T.L.V. and K.D. conducted all experiments and analyses. T.L.V., K.D., and K.G. wrote the manuscript.

Figures

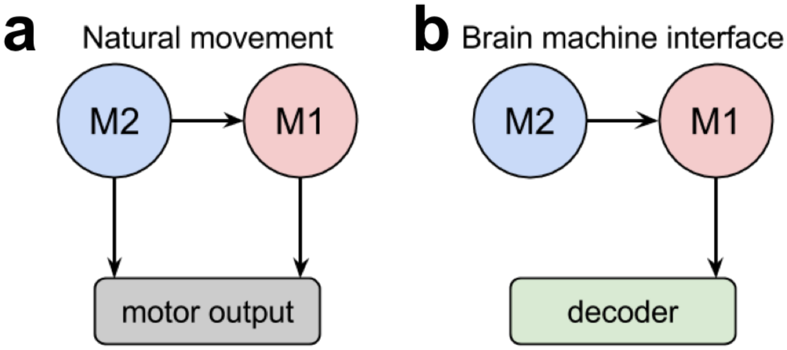


Figure 2.1. BMI Model.
(a) In natural movement, motor (M1) and premotor (M2) cortex both have subcortical output. (b) In BMI learning, M2 can only affect the output via its influence on M1.

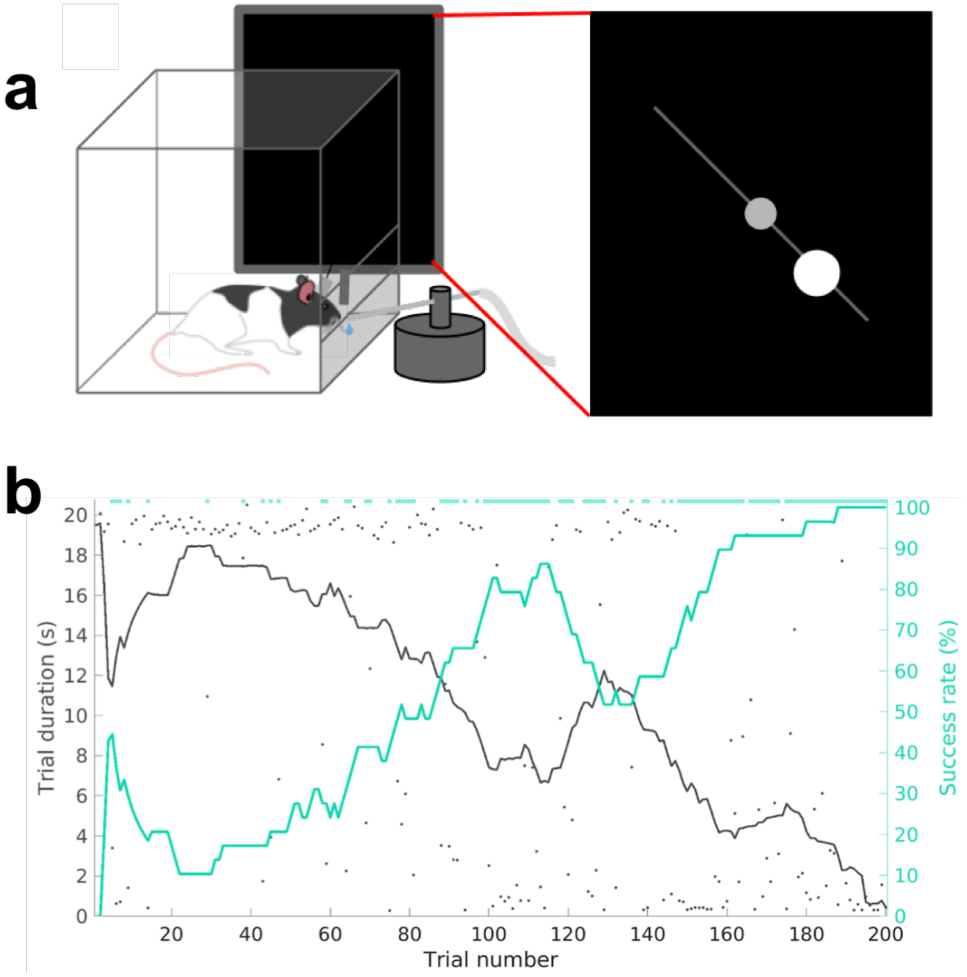


Figure 2.2. BMI task.
(a) The rat uses the BMI to bring the water spout to the reward position. In some animals, concurrent visual feedback was provided reflecting neural state. The grey circle represents the baseline state. The white circle represents the current brain state. Movement of the white circle is constrained to the diagonal line. (b) Example single-day learning curve for a robust learning session in one rat. The black line reflects a 30-trial average of trial duration. The green line reflects a 30-trial average of success rate.

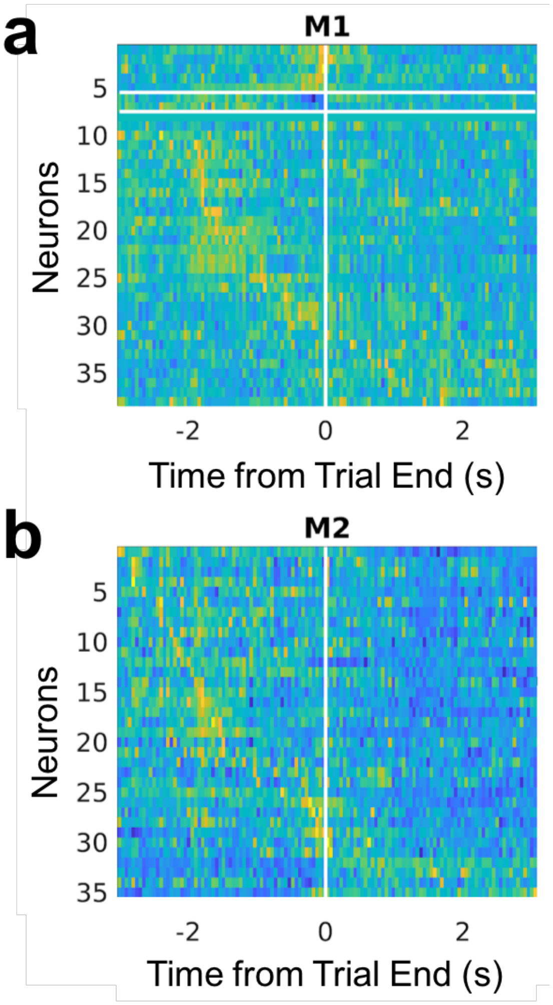


Figure 2.3. M2 is modulated by M1-BMI Learning
(a) Example z-score normalized peri-event time histograms (PETHs) of M1 neural activity aligned to trial end for a single learning session in one rat. Units in the top section are positively-modulated M1 direct units (pool 1). Units in the middle section are negatively-modulated M1 direct units (pool 2). Units in the lower section are M1 indirect units. (b) Same as (a) but for M2 neurons. No M2 neurons participate in the BMI decoder.

References

1. Hikosaka, O., Nakamura, K., Sakai, K. & Nakahara, H. Central mechanisms of motor skill learning. *Current Opinion in Neurobiology* **12**, 217–222 (2002).

2. Fetz, E. E. Operant conditioning of cortical unit activity. *Science* **163**, 955–958 (1969).

3. Neely, R. M., Koralek, A. C., Athalye, V. R., Costa, R. M. & Carmena, J. M. Volitional Modulation of Primary Visual Cortex Activity Requires the Basal Ganglia. *Neuron* **97**, 1356-1368.e4 (2018).

4. Athalye, V. R., Ganguly, K., Costa, R. M. & Carmena, J. M. Emergence of Coordinated Neural Dynamics Underlies Neuroprosthetic Learning and Skillful Control. *Neuron* **93**, 955-970.e5 (2017).

5. Athalye, V. R., Santos, F. J., Carmena, J. M. & Costa, R. M. Evidence for a neural law of effect. *Science* **359**, 1024–1029 (2018).

6. Ganguly, K., Dimitrov, D. F., Wallis, J. D. & Carmena, J. M. Reversible large-scale modification of cortical networks during neuroprosthetic control. *Nature Neuroscience* **14**, 662–667 (2011).

7. Gulati, T., Ramanathan, D. S., Wong, C. C. & Ganguly, K. Reactivation of emergent task-related ensembles during slow-wave sleep after neuroprosthetic learning. *Nat Neurosci* **17**, 1107–1113 (2014).

8. Sakellaridi, S. *et al.* Intrinsic Variable Learning for Brain-Machine Interface Control by Human Anterior Intraparietal Cortex. *Neuron* **102**, 694-705.e3 (2019).

9. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

10. Golub, M. D. *et al.* Learning by neural reassociation. *Nature Neuroscience* (2018). doi:10.1038/s41593-018-0095-3

11. Zhou, X., Tien, R. N., Ravikumar, S. & Chase, S. M. Distinct Types of Neural Reorganization During Long-Term Learning. *Journal of Neurophysiology* (2019). doi:10.1152/jn.00466.2018

12. Gulati, T., Guo, L., Ramanathan, D. S., Bodepudi, A. & Ganguly, K. Neural reactivations during sleep determine network credit assignment. *Nat Neurosci* **advance online publication**, (2017).

13. Koralek, A. C., Jin, X., Long Ii, J. D., Costa, R. M. & Carmena, J. M. Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills. *Nature* **483**, 331–335 (2012).

14. Hikosaka, O. *et al.* Parallel neural networks for learning sequential procedures. *Trends in Neurosciences* **22**, 464–471 (1999).

15. Makino, H. *et al.* Transformation of Cortex-wide Emergent Properties during Motor Learning. *Neuron* **94**, 880-890.e8 (2017).

16. Perich, M. G., Gallego, J. A. & Miller, L. E. A Neural Population Mechanism for Rapid Learning. *Neuron* **100**, 964-976.e7 (2018).

17. Rothwell, P. E. *et al.* Input- and Output-Specific Regulation of Serial Order Performance by Corticostriatal Circuits. *Neuron* **88**, 345–356 (2015).

18. Rouiller, E. M., Moret, V. & Liang, F. Comparison of the Connectional Properties of the Two Forelimb Areas of the Rat Sensorimotor Cortex: Support for the

Presence of a Premotor or Supplementary Motor Cortical Area. *Somatosensory & Motor Research* **10**, 269–289 (1993).

19. Hyland, B. Neural activity related to reaching and grasping in rostral and caudal regions of rat motor cortex. *Behavioural Brain Research* **94**, 255–269 (1998).

20. Saiki, A. *et al.* Different Modulation of Common Motor Information in Rat Primary and Secondary Motor Cortices. *PLoS ONE* **9**, e98662 (2014).

21. Hira, R. *et al.* In vivo optogenetic tracing of functional corticocortical connections between motor forelimb areas. *Front Neural Circuits* **7**, (2013).

22. Otchy, T. M. *et al.* Acute off-target effects of neural circuit manipulations. *Nature* **advance online publication**, (2015).

23. Jazayeri, M. & Afraz, A. Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017).

24. Ramanathan, D. S., Conner, J. M., Anilkumar, A. A. & Tuszynski, M. H. Cholinergic systems are essential for late-stage maturation and refinement of motor cortical circuits. *Journal of Neurophysiology* **113**, 1585–1597 (2015).

25. Gharbawie, O. A., Karl, J. M. & Whishaw, I. Q. Recovery of skilled reaching following motor cortex stroke: do residual corticofugal fibers mediate compensatory recovery? *European Journal of Neuroscience* **26**, 3309–3327 (2007).

26. Brown, A. R. & Teskey, G. C. Motor Cortex Is Functionally Organized as a Set of Spatially Distinct Representations for Complex Movements. *Journal of Neuroscience* **34**, 13574–13585 (2014).

27. Morandell, K. & Huber, D. The role of forelimb motor cortex areas in goal directed action in mice. *Scientific Reports* **7**, 15759 (2017).

28. Rothschild, G., Eban, E. & Frank, L. M. A cortical-hippocampal-cortical loop of information processing during memory consolidation. *Nat Neurosci* **advance online publication**, (2016).

29. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience* **36**, 337–359 (2013).

30. Kao, J. C. *et al.* Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nature Communications* **6**, 7759 (2015).

31. Schroeder, K. E. & Chestek, C. A. Intracortical Brain-Machine Interfaces Advance Sensorimotor Neuroscience. *Front. Neurosci.* **10**, (2016).

32. Cao, V. Y. *et al.* Motor Learning Consolidates Arc-Expressing Neuronal Ensembles in Secondary Motor Cortex. *Neuron* (2015). doi:10.1016/j.neuron.2015.05.022

33. Chuong, A. S. *et al.* Noninvasive optical inhibition with a red-shifted microbial rhodopsin. *Nat Neurosci* **17**, 1123–1129 (2014).

34. Tanji, J. Sequential Organization of Multiple Movements: Involvement of Cortical Motor Areas. *Annual Review of Neuroscience* **24**, 631–651 (2001).

35. Halsband, U., Ito, N., Tanji, J. & Freund, H.-J. The role of premotor cortex and the supplementary motor area in the temporal control of movement in man. *Brain* **116**, 243–266 (1993).

36. Hatsopoulos, N. G. & Amit, Y. Synthesizing complex movement fragment representations from motor cortical ensembles. *Journal of Physiology-Paris* **106**, 112–119 (2012).

37. Hatsopoulos, N. G., Xu, Q. & Amit, Y. Encoding of Movement Fragments in the Motor Cortex. *J. Neurosci.* **27**, 5105–5114 (2007).

38. Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F. & Fee, M. S. Growth and splitting of neural sequences in songbird vocal development. *Nature* **528**, 352–357 (2015).

39. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience* **17**, 440 (2014).

40. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

41. Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I. & Shenoy, K. V. Cortical Preparatory Activity: Representation of Movement or First Cog in a Dynamical Machine? *Neuron* **68**, 387–400 (2010).

42. Chung, J. E. *et al.* A Fully Automated Approach to Spike Sorting. *Neuron* **95**, 1381-1394.e6 (2017).

Chapter 3: Closed-loop Deep Brain Stimulation for Refractory Chronic Pain

Authors: P. Shirvalkar [1,3]‡ , T. L. Veuthey [2]‡, H. Dawes [3], E. Chang [3]

[1]Departments of Neurology and Anesthesiology (Pain Management Division), University of California San Francisco, San Francisco, CA, USA

[2]Neuroscience Graduate Program, University of California San Francisco, San Francisco, CA, USA

[3]Laboratory of Edward Chang, Center for Neural Engineering and Prostheses, Department of Neurosurgery, University of California San Francisco, San Francisco, CA, USA

‡ These authors contributed equally to this work.

Article Type: Hypothesis and Theory

Keywords: Closed-loop, Deep Brain Stimulation, Neuropathic Pain, Chronic Pain, Control Theory, Somatosensory, Affective, Cognitive

Abstract

Pain is a subjective experience that alerts an individual to actual or potential tissue damage. Through mechanisms that are still unclear, normal physiological pain can lose its adaptive value and evolve into pathological chronic neuropathic pain. Chronic pain is a multifaceted experience that can be understood in terms of somatosensory, affective, and cognitive dimensions, each with associated symptoms and neural signals. While there have been many attempts to treat chronic pain, in this article we will argue that closed-loop deep brain stimulation (DBS) offers an urgent and promising route for treatment. Contemporary DBS trials for chronic pain use 'open-loop' approaches in which tonic stimulation is delivered with fixed parameters to a single brain region. The impact of key variables such as the target brain region and the stimulation waveform is unclear, and long-term efficacy has mixed results. We hypothesize that chronic pain is due to abnormal synchronization between brain networks encoding the somatosensory, affective and cognitive dimensions of pain, and that multisite, closed-loop DBS provides an intuitive mechanism for disrupting that synchrony. By (1) identifying biomarkers of the subjective pain experience and (2) integrating these signals into a state-space representation of pain, we can create a predictive model of each patient's pain experience. Then, by establishing how stimulation in different brain regions influences individual neural signals, we can design real-time, closed-loop therapies tailored to each patient. While chronic pain is a complex disorder that has eluded modern therapies, rich historical data and state-of-the-art technology can now be used to develop a promising treatment.

Introduction

Chronic pain is a major healthcare problem, and estimates by the CDC suggest that it affects more people in the US than heart disease, diabetes and cancer combined [1]. Central neuropathic pain, defined by the International Association for the Study of Pain as pain originating from a lesion of the brain or spinal cord, is often refractory to treatments [2]. Common pharmacological therapies have marginal analgesic benefit, and so far, modern neuromodulation therapies such as spinal cord or deep brain stimulation have had limited efficacy over time. Currently, these therapies offer a one-size-fits-all approach that is not optimized for individual neural signatures of pain. However, we believe that central pain syndromes are particularly good candidate conditions for personalized medicine. Each patient's pain is a multifaceted experience that can be understood in terms of somatosensory, affective, and cognitive dimensions, each correlated with activity in different brain regions [3,4](Figure 3.1). We hypothesize that providing enduring analgesia will be best achieved by identifying patients' unique neurophysiological biomarkers of pain perception across multiple brain regions and providing tailored, feedback-controlled deep brain stimulation across those target regions. Importantly, we acknowledge that we seek not to abolish all pain perception per se, as pain may serve an adaptive role to averting tissue injury. In this article, we outline prior approaches to DBS for chronic pain, an approach to identifying neural biomarkers of pain, and propose strategies to develop a framework for closed-loop DBS based on control theory and state-space paradigms.

Hypothesis

A brief history of DBS for pain

Chronic pain has been conceptualized as a multidimensional process for many decades. Opioids, one of the most common therapies for chronic pain, incidentally provide relief for somatosensory, affective, and cognitive aspects of pain and target top down modulation of pain sensation [5–7]. However, most neuromodulatory therapies such as transcranial magnetic stimulation (TMS) and DBS still focus on a single facet of pain, originally targeting either somatosensory networks or more recently targeting affective regions. These therapies and their outcomes provide insight into the potential and limitation of addressing centralized pain syndromes as a single-modality pathology.

DBS for Somatosensory Pain Symptoms

Early efforts at targeting DBS for pain focused on modulating signals in somatosensory networks. Initial inspiration to target these brain regions was inspired by Dejérine and Roussy's descriptions of post-stroke pain syndrome in patients with thalamic infarcts involving the spinothalamic pathway [8]. In an attempt to silence aberrant activity in somatosensory pathways, patients underwent ablations of various segments along the spinothalamic tract and the dorsal thalamus [9,10]. Eventually, direct electrical stimulation of the dorsal column [11], internal capsule [12] and sensory thalamus [13] provided a reversible alternative to ablation.

Based on results from intraoperative microstimulation in humans, several groups designed studies to disrupt neural signals of different nodes in the somatosensory/nociceptive network. Since 1969, small case series targeting DBS to the ventral (or caudal) thalamus (vT), internal capsule, and periventricular / periaqueductal grey (PVG/PAG) were conducted with efficacy rates ranging from 23-59% [13–15]. To extend these case series, Medtronic conducted two large, multicenter, randomized controlled trials in the early 1990s for a heterogeneous group of chronic pain conditions [16]. All patients were implanted with bilateral electrodes targeted the vT and PAG. These trials established the primary endpoint still used by most modern chronic pain trials: >50% reduction of the pain visual analog score (VAS) at one year. However, they were aborted in the 1990s, largely due to poor enrollment and participant attrition. Around the same time, the FDA granted Medtronic approval of DBS for Parkinson's Disease (PD) and essential tremor and Medtronic never sought market approval for pain indications. Common criticisms of Medtronic's DBS trials for chronic pain include 1) poor patient selection due to wide heterogeneity of pain etiologies (i.e. nociceptive pain, neuropathic pain, thalamic pain, visceral pain, brachial plexus avulsion, unspecified etc.), 2) a minority of purely neuropathic pain syndromes (~30%) and 3) lack of appropriate patient follow up. This study used fixed, tonic stimulation parameters ranging from 100-130 Hz which were manually optimized at the start of the study for each patient. It remained unclear exactly how electrical stimulation affected targeted regions, but long-term pain relief waned likely due to adaptation of the nervous system to continuous stimulation and the development of tolerance. Despite this lack of mechanistic clarity, DBS became a compelling experimental therapy because it is still preferable to permanent ablation or resection of brain tissue which has low analgesic efficacy.

Early attempts to stimulate the somatosensory cortex directly failed to provide pain relief [17]. Instead, stimulation of the adjacent motor cortex with arrays of electrodes has been successfully used to treat pain syndromes such as pelvic pain [18], trigeminal neuralgia [19] and phantom limb pain [20], presumably by providing feedback inhibition of S1 inputs [17]. Efficacy rates of motor cortex stimulation range from 40-60% but significant long-term studies are lacking.

DBS for Affective Pain Symptoms

Based on animal studies implicating limbic system structures in emotional experience and expression [21,22], early brain surgery for chronic pain involved anterior cingulotomy to alleviate pain. Case studies of these patients described individuals with intact somatosensation, but who seemed to lack “emotional tension” [23,24] and lacked “emotional reactivity” to pain stimuli [25] without being emotionally blunted.

The earliest reports of DBS induced analgesia were actually serendipitous findings from stimulation of septal nuclei in patients with psychiatric disorders in the 1950’s [15]. These findings were not followed up until the 1960’s, when Lewin and Whitty performed intraoperative stimulation of the cingulate cortex which produced transient analgesia.

Modulating the affective component of pain has reflected a paradigm shift for DBS in the 21st century. Recent studies measuring cerebral blood flow with positron emission tomography (PET) or functional magnetic resonance imaging (fMRI) have specifically identified the dorsal anterior cingulate (dACC), insula, and dorsolateral prefrontal cortex

(DLPFC) as key substrates underlying subjective pain experience [26,27] of which the ACC may be specific to the affective component of pain [28]. Animal studies have further corroborated this evidence by demonstrating a causal role for ACC neurons in mediating the ‘aversiveness’ of nociceptive stimuli. Fields and colleagues demonstrated that destructive lesions of the rostral ACC reduce learned conditioned pain preference in a rat pain assay [29]. Injecting an excitatory amino acid into the ACC, even in the absence of a noxious pain stimulus, actually increases conditioned place preference, suggesting that the ACC is both necessary and sufficient for learning the ‘unpleasantness’ associated with pain stimuli [30].

Two cases of ACC stimulation for spinal cord injury have shown therapeutic promise [31], and another recent study demonstrated that stimulation of the anterior midcingulate cortex produced an attitude of resilience and ‘will to persevere [32].’ The first human clinical trial using open-loop DBS in ACC for chronic pain showed a significant decrease in pain ratings (Visual Analog Score) at one year with enduring relief at a two year time point [33,34]. A recent attempt to modulate the affective dimension of pain with DBS targeting the ventral striatum / anterior limb of the internal capsule for post stroke pain did not show improvement in pain scores, but did enhance measures of mood further implicating basal forebrain regions in distributed pain circuits [35].

Limitations to Current Approaches

Current clinical paradigms for DBS are all ‘open-loop’ systems, in which tonic stimulation is continuously applied to a single brain region. Constraints on electrode location and stimulation parameters limit the efficacy of open-loop DBS.

Anatomical limitations

By restricting stimulation to one brain region, traditional DBS fails to account for the fact that the hallmark of pain is not based on strong signals in any *single* one of the three components of pain (somatosensory, affective, and cognitive), but a confluence of signals in all three (Figure 3.1). We hypothesize that chronic pain is due to abnormal synchronization between brain networks encoding these three dimensions of pain. Consequently, effective pain relief is unlikely to be achieved by blunting a single component; instead, it will be more effective to decouple and modulate each of them through multisite stimulation. Below, we propose the following candidate brain regions as appropriate targets to test our hypothesis: primary somatosensory cortex (S1, somatosensory), dorsal anterior cingulate cortex (dACC, affective), and orbitofrontal cortex (OFC, affective and cognitive) (Figure 3.3).

Stimulation limitations

By restricting stimulation to fixed parameters, an open-loop strategy cannot take into account the fact that pain for a single patient comes in many forms. While some instances of pain are evoked by sensory stimuli, spontaneous and constant pain states are also influenced by mood and attention [6,36]. Based on personal pain symptoms, abnormal somatosensory signals will need to be modulated to different degrees than affective and cognitive signals in a time varying manner. Currently, stimulation parameters are tediously optimized by a healthcare provider by systematically changing variables such as pulse width, frequency and amplitude to find the settings that best provide a desired effect. Changes are made on the timescale of patient visits. Ideally, adaptive stimulation would change in real-time to match the dynamic changes in a patient's pain state.

Temporal limitations

Tonic, open-loop stimulation also does not account for the dynamic nature of pain or adaptation of the brain over time. Loss of therapy often occurs over months to years due to changing impedance of electrodes and development of scar tissue around contact sites. A feedback driven stimulation paradigm would ideally account for such changes and adjust the contact site or parameters of stimulation appropriately. Closed-loop DBS provides flexible solutions to limitations of open-loop approaches. Below, we describe a theoretical framework for design of a feedback controlled (closed-loop) DBS system to address the multiple dimensions of chronic pain using state-space control theory.

Applying control theory to DBS for pain

Pain can be studied, understood, and treated through different levels of abstraction. Prescribing opioids inherently addresses pain as a chemical process. Here we will address pain as a network process. Through this lens we will analogize pain to a dysfunctional signal within an electrical network, which itself is limited to a few components within the central nervous system. In this analogy, managing pain can be addressed as a control systems problem, in which the brain is the component we are trying to regulate, and the DBS device is the control box. The availability of different control systems, particularly open-loop versus closed-loop devices, leads to different goals and approaches. However, no artificial system will be a full substitute for a healthy human pain system, which relies on access to widespread brain regions to provide pain control that is influenced by mood, social context, physical modality, emotional valence, attention and temporal structure. We suggest that both open-loop and closed-loop strategies should set realistic goals, such as identifying and preventing both constant and spontaneous pain states and/or giving patients more control over their pain treatment.

Mapping DBS onto a control framework

We would like to clearly map out the analogy between classic control schemas and pain control through external devices. Figure 3.2A shows the classic layout of a feedback driven control system, and Figure 3.2B shows how different components of DBS as a medical intervention map onto each role. The system in question is the brain itself, specifically the pain-related regions with pathological pain signals. The system output is an observable biomarker which we hypothesize as giving an accurate, relevant, and temporally appropriate view into the patient’s pain state. The sensor is any implanted recording electrodes (e.g.: microwire arrays, ECoG grids, EEG leads), which records neural signals. The reference signal is the desired version (pain-free) of the neural signal. A closed-loop device would compare the sensed neural signals to the reference signal (measured difference) and trigger the DBS device (controller) to appropriate corrective stimulation, with the assumption that stimulation can control the internal state of the patient.

An *open-loop* system would be limited to the components in the red box (Figure 3.2B). Since there is no sensor, the output of this system is the patient’s self-report of pain. The healthcare provider compares this self-report to a reference, pain-free state and can adjust DBS stimulation parameters as needed. The timescale of updates is clinic visits, and there is no view into underlying neural signals related to pain. A *closed-loop* system (minimally defined as any system with in which stimulation is based on a sensor readout) gives access to neural signals that are interpreted as real-time proxies for the patient’s internal pain state. This readout is fed back and compared to a reference neural signal. Based on the difference between these signals, a controller makes responsive, real-time adjustments to stimulation parameters. It is the hope that closed-loop paradigms will improve outcomes and reduce side-effects compared to open-loop paradigms.

State Space Models

State-space representations are used in control engineering to model systems with multiple inputs, multiple outputs and latent state variables which can be used to represent dynamic sequences of brain states [37,38]. Neural state spaces representations can consist of a number of time dependent input variables, such as firing rates from neurons or local field potential (LFP) power time series from multiple recording channels. If the number of variables (i.e. neurons or electrode contacts) is very large, it is useful to first reduce the dimensionality of the data to a set of orthogonal dimensions that describes the phenomena of interest with fewer variables [39]. This dimensionality reduction is commonly done with tools such as principal components or factor analysis which can help to identify *latent* variables that define a new coordinate system. Temporal evolution of the neural signal through this coordinate system can be interpreted as ‘neural trajectories.’

Recently, state-space representations have been used to understand the evolution of neural signals from motor cortex during reaching tasks [40,41]. The relationships between external triggers (visual reach target onset, go cue), internal state (movement preparation), conscious experience (anticipation), and behavior (movement onset) are intuitive for motor processes, and we argue that applying state-space analysis to pain dynamics may be similarly useful. While dynamical systems analysis of movement has so far mostly relied on single-neuron signals, there are also ample reports of using LFP from motor cortex to decode movements and screen cursor location [42–45]. Because shifts in pain state are slow, multiregional neural phenomena, we predict that LFP changes across multiple brain regions will provide a temporally appropriate neural report of pain state fluctuations. Multivariate data such as LFPs from multiple brain regions can be represented in a 'state-space' for pain (Figure 3.4). These are particularly appropriate for analyzing multidimensional phenomena like dimensions of pain. In the next section, we will outline the specific nature of the neural signals which can be interpreted as biomarkers of internal pain states.

Local Field Potentials Are the Most Tractable Signal for Identifying Biomarkers for Closed-Loop DBS

Candidate neurophysiological biomarkers for chronic pain can be derived from three types of signal: single action potentials, local field potentials (LFP) within specific frequency bands, and blood oxygen level dependent (BOLD) signals.

Single action potentials are the neural signal with the highest temporal and spatial resolution. However, action potentials collected from chronically implanted tungsten or silicone probes are unstable due to probe drift and sensitivity to behavioral context (i.e.: sensory stimulation, arousal state, etc). Single action potentials from S1 and ACC were used in a rodent model of acute thermal pain to decode a pain state defined through use of a Hidden Markov Model [46]. In this experiment, signals from the population of single neurons used to computed baseline and pain states were not stable over even a few trials, making the chronic computation of a pain state untenable. Assuming that recorded action potentials from human patients would experience similar instability, chronic biomarkers based on these signals are not tractable. A potential work-around would be to calculate biomarkers based on dynamics from population neural firing combined with high frequency local field potential, a promising strategy used in human brain-machine interfaces [47].

Local field potentials represent aggregate population subthreshold activity among a spatially localized population of neurons [48]. While the term LFP usually refers to signals captured by implanted depth electrodes or cortical electrodes, LFP is thought to reflect brain oscillations similar to those captured by intracranial electroencephalography (iEEG) and magnetoencephalography (MEG). Previous attempts at decoding subjective pain intensity with resting state EEG [49] or MEG [50] have used time-frequency representations of brain oscillations with high accuracy, supporting the feasibility of using LFP to define a pain state. Additionally, LFP signals are 1) easier to record than spikes or evoked potentials over single trials 2) often highly reproducible within an individual and 3) can be examined by well-developed signal analysis tools [51]. Previous studies of Parkinson's disease have successfully used LFP from depth electrodes and cortical strips to define biomarkers for tremor and dyskinesia over many days/months, providing support for the stability and longevity of this signal type [52]. In a closed-loop DBS trial for chronic pain, the healthcare team could record from multiple brain regions simultaneously to track changes in the multiple parallel dimensions of pain: somatosensory (S1, vT, insula), affective (ACC, medial thalamus, and striatum) and cognitive (PFC, OFC, insula).

Finally, several studies have used blood-oxygen-level dependent contrast imaging (BOLD signals) to detect and define pain states in fMRI research [27,53,54]. BOLD signal can reflect brain activity at a spatial resolution under 1 mm and at a temporal resolution of a few seconds [55], providing excellent whole-brain localization and temporal tracking of neural activity correlating with pain states. Unfortunately, these signals are not available in the ambulatory setting, prohibiting their use in chronic patient therapy. Also, current closed-loop DBS probes are not MRI compatible, and it is unclear whether these probes would cause signal artifacts once they are implanted. However, asking patients to complete a pre-implantation fMRI study to capture neural signals correlated with spontaneous and evoked pain would be extremely useful to direct patient-tailored anatomic targeting of the probe implant. Ideally, the healthcare team would capture simultaneous fMRI and EEG signals, which could also inform the initial search for LFP-based biomarkers (assuming that LFP signals provide a local view of neural signals more grossly captured in EEG) [56].

Computing a pain state from regional biomarkers

Pain is a multi-faceted process that can be broken down into somatosensory, affective, and cognitive components [4]. Each component can be associated with distinct symptoms and brain regions. Importantly, information processing for each component is not fully segregated, but instead involves activity in overlapping neural pathways. Currently, constellations of somatosensory, affective, and cognitive signs and symptoms are integrated by healthcare providers to characterize each patient's pain state. For example, two patients with back pain might have different locations and intensities of pain and might also be more or less bothered and distracted by that pain. Ideally, a complete description of a patient's pain state contains all of these components.

Similarly, neural recordings from different brain regions could be integrated to provide a multidimensional neural signature of a patient's pain state (Figure 3.3). Through this neural report, closed-loop brain stimulation becomes a tractable strategy for addressing dynamic pathological brain states. Based on real-time representations of a patient's pain within a neural state space, a closed-loop system can stimulate different brain regions to normalize different components of pain. Such a real-time representation of pain requires accurate and reliable detection of neural biomarkers for somatosensory, affective, and cognitive components of pain. Like patient-reported symptoms of pain, these biomarkers can be thought of as the observable markers of the pain state.

We argue that using LFP signals from three brain regions – S1, dACC, and OFC— could be used to calculate multidimensional, patient-specific pain states (Figure 3.3).

(While we believe these brain regions are critical sites for detecting pain signals there are other valuable regions that have been omitted for clarity in Figure 3.3). Each patient's biomarkers will need to be determined empirically, but based on prior literature (elaborated below), we suggest using high gamma power in S1, high gamma and low alpha power in dACC, and low alpha power in OFC as starting points. Patients' pain states endogenously fluctuate through the day, with higher pain experienced at some time points (i.e. Mornings, evenings) and lower pain states expected during periods of rest, sleep or after medication. To identify biomarkers of pain-states, we suggest sampling neural recordings and coincident pain scores during a wide range of naturally fluctuating chronic pain states in the ambulatory setting. Once neural data are collected, they can be transformed to a time-frequency representation to calculate power spectral density. Then, spectral density values within bands of interest (theta, alpha, gamma, etc.) should be used as independent variables to predict pain scores (dependent variables) (see section 6.1). The most predictive variables or combination thereof would serve as optimal biomarkers from each brain region.

Somatosensory Signals

The somatosensory-discriminatory component of pain encompasses the intensity, location and duration of a noxious stimulus ('what', 'where' and 'when'). This component of pain has been the most widely studied and is often modeled with transient acute painful stimuli such as electric shock or a phasic thermal or laser pain stimulus lasting a few seconds. As a first step towards decoding chronic pain states, it may be helpful to study decoding of acute pain stimuli, though it is critical to distinguish biomarkers of pain perception from mere pre-perceptual stimulus processing. Human functional imaging data point to a widely distributed neural network that is activated by acute experimental pain perception including the primary and secondary somatosensory cortex (S1 and S2), insula, ACC, PFC and thalamus [26,27,57]. However, not all signals can strictly be interpreted to represent somatosensory perception.

A recent study using magnetoencephalography (MEG) to identify neural correlates of cutaneous laser evoked pain in healthy human subjects showed an increase in gamma band amplitude (65-90 Hz) in the contralateral S1 at 200-400 msec post stimulus onset [58]. This gamma increase was predictive of subjective pain intensity and persisted when controlling for stimulus salience or attentional (cognitive) effects by presenting a stimulus repeatedly [59]. Therefore, gamma activity may represent pain perception and not just stimulus processing. However, many of these studies lack non-painful control stimuli, making it possible that gamma activity reflects somatosensation more generally.

Baseline EEG recordings of patients with chronic neuropathic pain show increased theta (4-10 Hz), alpha (12-20 Hz) and beta (20-30 Hz) band power in the insula, frontal cortices, and anterior cingulate [60,61] which may reflect multiple pain dimensions. There is a further trend towards global slowing with lower peak alpha and theta frequencies in patients with neuropathic or thermal pain [62] which is not seen in nociceptive pain [63]. Further, suppression of alpha band oscillations is commonly reported after acute pain stimuli (further discussion below,[64]). Together these data point to band-limited power changes in S1, insula and thalamus as candidate biomarkers for somatosensory-discriminative pain perception.

Given the pragmatic need to select a single somatosensory region from which to derive pain signals, we suggest recording in S1 rather than vT because cortical regions have higher amplitude signals and may be more reliable over time. While optimal somatosensory biomarkers are best determined empirically for each patient, filtered high gamma power has been a consistent marker in several studies and provides a reasonable starting point as a feedback-control signal for closed-loop DBS. After correlating the relationship of gamma power to patient-reported pain scores, values for gamma power that reliably distinguish high pain states from low pain states can be used to define a high pain-state detection threshold. Then, threshold crossing of real-time gamma power, in combination with other regional biomarkers, can be used to automatically activate analgesic stimulation as needed (see section 6.1 for details).

Affective Signals

The affective dimension involves the ‘unpleasantness’ of a stimulus, and is tied to motivation to rid the pain, changes in mood and anxiety and the degree of suffering [36]. Brain regions underlying affective encoding were identified using positron emission tomography (PET) in subjects undergoing hypnosis to selectively reduce the ‘unpleasantness’ of acutely painful stimuli [28]. While individuals still felt similar intensity of pain stimuli under hypnotic suggestion, they were not bothered by these stimuli; they showed reduced activation of the ACC (but not S1) which was linearly related to pain unpleasantness. The role of the rostral ACC in the affective dimension of pain is also corroborated by a recent large meta-analysis of over 10,000 functional MRI datasets [53] and animal studies that support the role of the medial ACC in transition from acute to chronic pain which has a larger affective component [65].

Tonic pain stimuli lasting longer than 10 minutes are likely closer to modeling chronic pain states, and engage distinct brain regions from acute pain stimuli [66]. EEG recordings in humans point to increased amplitude of gamma band oscillations in the cingulate and medial prefrontal cortex after tonic pain stimuli [67,68].

Animal studies also help to identify brain regions and candidate signals that may serve as affective biomarkers of pain perception. In a study recording single spikes from S1 and ACC of rats, a state space model was used to identify neuronal codes underlying acute painful thermal stimuli that produce a paw withdrawal reflex [46]. One key insight from this study was that population spiking activity from S1 provides better sensitivity for acute pain prediction, while activity from ACC provides better specificity suggesting that a subset of neurons in ACC encode pain information. Simultaneous single neuron recordings in mice in S1, vT, ACC and mediodorsal thalamus (MD) show temporal and lateralized segregation of encoding of noxious stimuli [69]. While S1 and vT cells predominantly fired early and contralateral to the pain stimulus, MD and ACC cells had long lasting firing which correlates with the longer time course of pain related anxiety or mood. These data further support the role of the MD thalamus or ACC in affective pain processing. Because cortical signals provide easier surgical access and higher amplitude LFP signals, ACC would be a reasonable initial brain target.

Cognitive Signals

Cognitive aspects of the pain experience involve implementing successful coping strategies, pain anticipation/ expectations and behaviors related to attention and distraction [36]. Increased attention to a painful stimulus will increase the perceived intensity of pain without altering its unpleasantness; distraction from pain can be analgesic. Further, pain itself often interferes with attentional processes, making causal inference of the role of attention difficult. Cognitive strategies that reduce pain perception such as distraction increase the amplitude of EEG activity in the DLPFC, orbitofrontal cortex (OFC) and caudal ACC shortly after a pain stimulus [70]. Modulation of the alpha rhythm is widely associated with the cognitive component of pain. Intracranial recordings in epilepsy patients suggests that increased attention towards a painful stimulus is correlated with alpha and beta band activity in the medial PFC and parasylvian regions that exert a causal influence over S1; this relationship is the opposite with distraction [71]. Similar alpha coherence between PFC and S1 is seen in ECoGs during pain anticipation [72]. Further, the amplitude of frontocentral alpha correlates with subjective expectation of pain relief induced by placebo [73]. These observations support the role of perisylvian regions such as PFC and OFC, and alpha band oscillations in the cognitive dimension of pain.

Oscillations before the onset of pain can shape the experience of pain and may serve as a context dependent biomarker of cognitive control over pain. Two recent studies show that the amplitude of pre-stimulus alpha oscillations (12-20 Hz) over somatosensory cortex is inversely correlated with pain perception [74,75]. However, multiple other studies report changes in alpha power of the PFC with attention and perception of non-painful stimuli, confounding general interpretation of this effect. Functional imaging and EEG studies further point to functional connectivity between the PFC, anterior insula and temporoparietal junction that form a 'salience network' that underlies cognitive control over pain [76].

Based on the available literature, OFC would be a reasonable initial target to identify putative pain biomarkers of the cognitive-evaluative dimension.

Multidimensional biomarkers for chronic pain

By simultaneously recording intracranial LFPs in multiple brain regions, it may be possible to identify biomarkers for unique pain states (spontaneous pain flare, evoked pain, baseline pain) that are more sensitive and specific than any single brain region can provide. Further, frequency band-limited activity between these brain regions is interpreted to reflect the flow of information [66,72,77], more accurate prediction of pain states may result from calculating phase coherence or amplitude co-modulation between each region's signal. Recent evidence suggests phase or amplitude relationships *between* different frequency oscillations *within* a brain region may also be informative about information flow [60,78,79] as in a model of closed-loop DBS for Parkinson's Disease [52].

We argue that using LFP signals from three brain regions – S1, dACC, and OFC— could be used to calculate multidimensional, patient-specific pain states. While each patient's biomarkers will need to be determined empirically, based on prior literature (elaborated below), we suggest using high gamma power in S1, high gamma and low alpha power in dACC, and low alpha power in OFC as starting points. Patients' pain states endogenously fluctuate through the day, with higher pain experienced at some time points (i.e. Mornings, evenings) and lower pain states expected during periods of rest, sleep or after medication. To identify biomarkers of pain-states, we suggest a protocol that involves sampling neural recordings and coincident pain scores during a wide range of naturally fluctuating chronic pain states in the ambulatory setting. Once neural data are collected, they can be transformed to a time-frequency representation to calculate power spectral density. Then, spectral density values within bands of interest (theta, alpha, gamma, etc.) should be used as independent variables to predict pain scores (dependent variables) (see section 6.1). The most predictive variables or combination thereof would serve as optimal biomarkers from each brain region.

Ideally, a multidimensional biomarker (based in regions relevant to dimensions of pain) will define a pain 'landscape' that will distinguish pain states to be avoided from pain-free states that are desired (see Figure 3.4). In this theoretical framework, the next challenge is characterizing the dynamics of how brain activity in the above regions naturally enters and exits this pain state. As such, the boundaries of a pain state biomarker can be established by setting an appropriate threshold. The ideal goal of a closed-loop DBS paradigm is to prevent the onset of a pain state, rather than simply aborting it once it has commenced. By characterizing the causal consequences of different patterns of brain stimulation, we can determine the optimal stimulation parameters needed to avoid pain states, at multiple points in the landscape. Neural activity is adaptive, however, and this pain landscape may evolve over time making it difficult to define stable boundaries of pain-free states.

Computing a reference state

We can define a pain-free reference state with the same protocol (Section 4.4) used to define the boundaries of a high pain-state (Figure 3.4). The role of a reference state is to define the range of biomarker values which in turn will guide selection of a threshold to trigger stimulation. In practice, a ‘reference’ state would reflect any value of the biomarker below a defined threshold for high-pain (i.e. NRS> 7). In this view reference can simply be interpreted to mean ‘non-high pain state.’ Empirical data from chronic human recordings is needed to understand the stability of pain-state detection thresholds. Higher instability would require more frequent re-calculation in order to provide a meaningful contrast between the reference and pain states. Ideally, such a signal would be usefully stable on the order of months, but, but it may be reasonable to perform automated re-calibration monthly or weekly. Potential lapses in therapeutic stimulation can be identified by the patient who can trigger recalibration to update the model. This updating will entail definition of new pain-state thresholds and the selection of new stimulation parameters.

Alternatively, a reference state can be interpreted to mean a ‘low pain-state’ where numerical pain scores would be <3, for example. The possible utility of separately defining such a reference state has been suggested by a computational model for closed-loop control for to treat essential tremor in non-human primates [80]. In this model, investigators developed a closed-loop control system that automatically adjusted DBS stimulation amplitude based on the spectral content of simulated LFPs from a cohort of 100 neurons in the Vim thalamus. Optimal DBS output to suppress tremor replaced the tremor-related pathological LFP spectrum with LFP patterns similar to those simulated in a ‘reference’ tremor-free state. Similar approaches may help control stimulation amplitude in multiple brain regions based on expected ‘pain-free’ regional LFP spectra.

Modulating a pain state with different stimulation paradigms

Analogous to how biomarkers are selected to best delimit a pain state, stimulation parameters must be optimized to best control the pain state-space trajectory (Figure 3.4A). The pre-defined goals for stimulation control are to either abort or avoid pain states. The stimulation goal and parameter selection will depend on the control paradigm: open-loop, patient-triggered, sensor-triggered (on/off), or true closed-loop.

Open-Loop Stimulation

In an open-loop paradigm, the goal must be to avoid pain states because there is no sensor available to detect them (which would be necessary to abort them). Therefore, the stimulation parameters must be chosen to maintain the neural state in the pain-free zone (see Figure 3.4B). We hypothesize that this is best accomplished by consistently de-coupling the neural signals in each pain-related region. For example, leads in S1 and ACC could be alternately pulsed at high gamma frequencies to disrupt the ability of the two regions to develop pathological coherence. If only one stimulation site is available, we suggest targeting ACC rather than S1 or OFC, given recent promise in clinical trials [34]. Decoupling ACC from other regions might be accomplished by tonically inputting local entrainment signals that would block information flow about inappropriate pain. Similarly, local decoupling has been proposed as a hypothesized target for the treatment of hyperkinetic states in DBS for Parkinson's Disease [81].

Once the stimulation is turned on, the goal is indefinite avoidance of a pain state. However, onset of the therapeutic effect may take a few days, as continuous pain states fluctuate on the time course of days, and we expect the pain dynamics to have some 'inertia'. In a recent trial of open-loop ACC stimulation for chronic pain [33], there was a wash-in period of many days for any therapeutic effect.

Patient-Triggered Stimulation

In a patient-triggered paradigm, the goal is to abort pain states detected by the patient. Effective stimulation must be able to halt pain quickly, making the therapy more suitable to modulation of transient, breakthrough pain. Somatosensory signals are the best candidates for interruption in a single-region stimulation paradigm because we assume they have faster dynamics and often begins 'upstream' in the pain triggering process. We suggest targeting ventral thalamus or motor cortex (adjacent to S1) for single-region, patient-controlled gamma-frequency stimulation, based on previous partial success of these therapies [82,83]. Long-term tolerance to stimulation seen in previous vT trials might be prevented by limiting stimulation to brief, patient-triggered periods. However, because chronic pain can be also triggered by affective and cognitive events, such as stress and rumination, a somatosensory-only detection paradigm leaves patients vulnerable to breakthrough pain. Multi-region stimulation in S1 and ACC (or OFC) would aim to de-couple these regions, but the insidious time course of pain dynamics in ACC and OFC may belie optimal control of breakthrough pain. Prior work with 'preventative' devices for epilepsy had a 40% failure rate in preventing seizures, highlighting the limitations of an abortive strategy for neuromodulation [84]. Altogether avoiding entry into pain states requires online tracking of the pain state's ongoing dynamics.

Sensor-Triggered Stimulation

Instead of relying on external input, a fixed stimulation protocol can be triggered based on the detected position and/or trajectory of the state within in the neural manifold (Figure 3.4B). To make this possible, the device must include sensors that can detect relevant biomarkers and an algorithm to decode the pain state with a latency short enough to allow intervention. This is commonly referred to as adaptive DBS (aDBS). As we hypothesize that continuous pain states arise from maladaptive coherence between regions involved in in pain processing, multi-area coherence may be an ideal signal to track the underlying neural state. We propose tracking gamma coherence between S1, ACC and OFC. Preliminary recordings from pain elicited by somatosensory and cognitive events (i.e. touch, asking the patient to attend to their pain) would allow investigators to determine a threshold of gamma coherence to characterize the pain state. Thereafter, coherence values close to that threshold would trigger de-synchronized stimulation in each region in order to prevent further evolution of inter-regional coherence. Side-effects of S1-OFC stimulation are unknown, however. It is possible that pain dynamics may evolve too rapidly to be interrupted before a noticeable pain threshold is breached, leading to breakthrough pain. Decreasing the threshold for allowable coherence may address this shortfall. Overall, sensor-triggered stimulation is a reasonable staring point in the quest to develop new feedback-controlled paradigms. A promising alternative is to implement a closed-loop paradigm with the possibility to continuously manipulate underlying neural states.

Closed-Loop Stimulation

In a truly closed-loop system (as we define it), unique stimulation patterns are delivered based on the real-time predicted course of the pain trajectory. This is distinguished from sensor-triggered stimulation in that in a closed-loop protocol, the same coordinate location in a state space may trigger different stimulation patterns or update stimulation parameters (pulse width, frequency, amplitude) dependent on the history and context of the neural trajectory. For this to be possible, we must create a predictive model of the multidimensional pain state such that the future path of each trajectory can be determined based on history and the current state [40,85] (Figure 3.4A). There are several methods for producing such a model, including (1) modeling the state as a three dimensional flow field [86,87], or (2) creating a map outlining the probability of transitioning from any point in the field to every other point (i.e. Hidden Markov Models, [88]).

Based on the assumption that stimulation can control or influence neural trajectories related to pain, the model must additionally contain predictions about the effect of stimulation on the pain trajectory (Figure 3.4B). Typically, characterizing the input-output (IO) relationship between stimulation and neural state in DBS is a painstaking manual process whereby a clinician systematically varies stimulation parameters (pulse width, amplitude, frequency) and records consequent changes in the neural state [89,90]. A promising method to automate stimulation parameter optimization involves the use of a ‘binary noise’ modulated stimulation pattern whereby a range of parameters are stochastically sampled and used for stimulation [91]. With simultaneous neural recordings, one may use binary noise to define IO dynamics of a closed-loop DBS system more efficiently. However, both of these methods risk producing uninterpretable IO relationships if the timescale of stimulation parameter changes does not match the timescale of state space changes (e.g. long wash-in latency for therapeutic effect).

There are many pragmatic barriers to implementing a truly closed-loop system. First, only few devices with dual sensing and stimulating functions are approved for chronic implant in humans: NeuroPace RNS, while other devices are investigational only: Braingate system and Medtronic Activa PC+S device [92–94]. Because there are no chronic, invasive cortical recordings from candidate stimulation regions in patients with chronic pain, it is unclear whether the hypothesized biomarkers will provide sufficient observability of the internal pain state. Second, while computing multi-dimensional state spaces from neural data is routinely done to visualize offline data, implantable devices have not been optimized to perform these computations online. It is unclear what amount of computation will be viable to perform for continuous pain monitoring. Third, because there have been few long-term successes from small-scale trials of DBS for pain, it is unclear whether chronic pain states will be controllable via stimulation. Resolving this uncertainty will require a chronic, multi-site, sensing and stimulating device that allows for rapid exploration of a large range of stimulation parameters. Finally, one of the hopes for closed-loop stimulation is that it will allow for a reduction of current dosage (relative to continuous stimulation in open-loop paradigms), thereby increasing the device's battery life and reducing the side effects of unnecessary stimulation. However, optimizing stimulation based on battery life will require additional trade-offs, such as deciding on a pain threshold at which stimulation will be initiated, limiting duration of stimulation bouts to the minimum required for pain relief, and potentially sacrificing benefits of long-term stimulation, such as learned desynchronization of pain-related regions.

Conclusions

Pragmatic Considerations for a Closed-Loop DBS Protocol

Above we provide evidence that spectral power of oscillations within specific frequency bands (e.g. theta, alpha, gamma) shows changes in relevant brain regions that may predict low or high pain states. By recording theta, alpha and gamma oscillations from the LFP signal in S1, ACC and OFC during natural fluctuations in a patient's chronic pain state, we can compute spectral power density during periods of high pain states and so define a neural state space model for predicting chronic pain. For this purpose, the low pain state can be interpreted as a 'baseline' or reference state.

To compute a time-frequency representation of the raw LFP signal, we use a variant of the discrete Fourier Transform (DFT). There are multiple methods to implement a DFT-we suggest using the multitaper DFT implemented in the Chronux Toolbox for MATLAB, which reduces broadband bias [51]. To adequately sample pain states, we propose to use at-home, patient-triggered recordings be collected. Two data collection schemes can be used as needed 1) 60-second recordings can be scheduled at pre-set time points throughout the day (i.e. 8 am, 12 noon, 4 pm, 8 pm) or 2) activated by patients by pressing a button on their DBS programmer. Self-reporting of pain numerical rating scores can be done via an automated text-messaging system. For convenience, patients can be prompted up to 4 times per day to report pain scores and trigger recordings if pre-scheduled recordings are not set. Once a series of Pain scores spanning a wide range (at least 5 different values on the numerical rating score) are collected, putative biomarker features can be used (as independent variables) to predict high (>7/10) or low (<4/10) pain score (dependent variable).

One possible solution to predicting low vs high pain states is to use multivariate logistic regression using biomarker features as independent variables, and low / high pain state as the dichotomous independent variable to be predicted. For example, if the ACC signal shows increased gamma power, and OFC shows decreased theta power during high pain states compared to baseline in an individual patient, predictive value of ACC gamma and OFC theta would be established through a multivariate logistic regression to predict pain state. A classification table would be used to calculate the probability of false positives and negatives, and overall prediction accuracy. In depth methods for developing multivariate classifiers based on logistic regression have been presented previously [95], as has their personalized application to closed-loop DBS systems based on brain-state [96]. Using Receiver operating characteristic (ROC) curves, one could then calculate optimal threshold values for each biomarker such that real-time crossing of ACC gamma or OFC theta power above/below this threshold would activate stimulation. This scheme represents a sensor-triggered protocol which is a good first-step approximation to building a fully closed-loop system that would adjust stimulation amplitude or other parameters based on ongoing neural activity.

Solutions to developing fully-closed loop algorithms and optimizing stimulation based on biomarkers have been suggested by computational studies. Recent models have used LFP spectra (beta and gamma power) as feedback-control signals to provide efficient and selective target stimulation in Parkinson's Disease [97] and essential tremor [80]. Using finite element modelling, anatomical models from imaging data can be combined with electrical models to optimize how current is delivered from the DBS electrode [98]. Further, stimulation patterns derived from computational evolution models may provide more battery-efficient stimulation protocols that can augment energy savings afforded by closed-loop DBS [99]. While explicit models have not been reported for closed-loop control in chronic pain states, future studies will need to incorporate multiregional brain recording and stimulation in relevant areas to provide analgesic closed-loop DBS.

As of the writing of this paper, our group is currently enrolling patients for participation in a feasibility study to develop closed-loop DBS algorithms for chronic neuropathic pain (ClinicalTrials.gov ID# NCT03029884). This trial seeks to enroll 10 patients with refractory neuropathic pain syndromes over 2 years and aims to develop a personalized treatment for multiple pain disorders using the Medtronic Activa PC+S device.

Discussion

The current article makes several important assumptions to create a simple theoretical framework for implementing closed-loop DBS for chronic pain syndromes. First, to disentangle biomarkers related to the somatosensory, affective, and cognitive components of pain, we impose artificial distinctions between brain regions that underlie each dimension of pain. We do not actually believe that chronic pain can be divided into three segregated, independent components with corresponding brain regions Rather, coalitions of cells in specific brain regions provide overlapping and complementary information about pain states.

Second, we would like to acknowledge that DBS provides an artificial input that may drive neural signals into unnatural regions of the pain based state space [100]. We hypothesize that such an induced discrepancy with natural states leads to reduced efficacy and increased side-effects. One of the main potential benefits of closed-loop stimulation would be to modulate neural signals to stay within natural bounds of information processing as seen in endogenous pain-free epochs.

Finally, the proposed framework assumes that optimal biomarkers come from neural signals. However, there are many other correlates of pain that provide useful signals. For example, the RESTORE trial matches different spinal cord stimulation parameters with different patient body positions, determined from an implanted 3D accelerometer [101]. Adapting stimulation to time of day, medication timing, sleep metrics, and other external variables would also improve intervention efficacy. Broadly speaking, we acknowledge that open-loop paradigms still incorporate a form of feedback, but on the timescale of clinic visits. Ultimately, all stimulation protocols for pain including personalized closed-loop models are designed based on offline analysis by the healthcare team. The most important ‘biomarker’ relevant for determining efficacy will always be the patients’ self-report of pain.

Acknowledgements

We thank Howard Fields, Maryam Bijanzadeh, Kristin Sellers and Katherine Scangos for their comments on this manuscript. **Funding:** This study is funded by Department of Defense: Defense Advanced Research Projects Agency (DARPA) SUBNETS grant #P0063517 and National Defense Science and Engineering Grant Fellowship (NDSEG). **Conflicts of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Figures

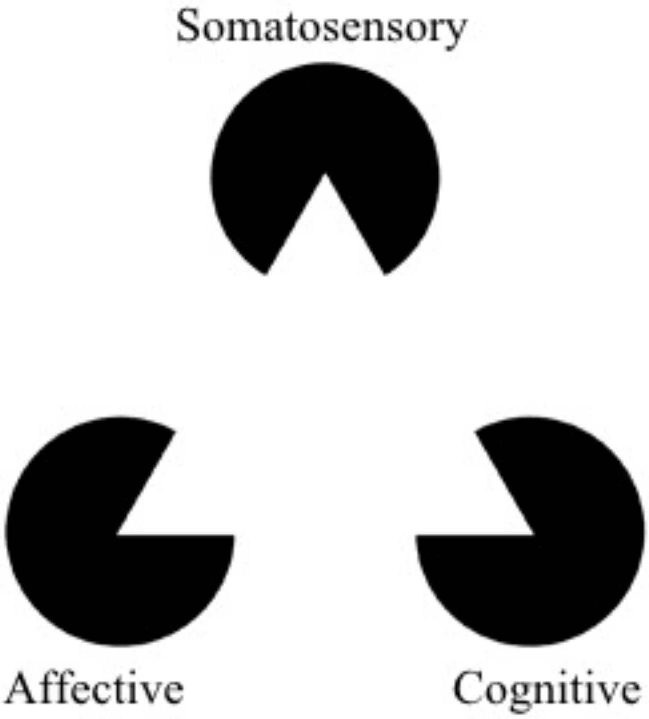


Figure 3.1. The Kanizsa triangle can be used to represent a multidimensional framework for pain.

Pain is an underlying state made apparent by three types of observable symptoms (somatosensory, affective, and cognitive). Therapies which selectively address a single facet of pain risk misinterpreting aspects of symptoms (the “shape” of the symptoms) outside of the context of the larger pathology. The optimal way to “break” the pain state might lie in modulation (or “re-orienting”) the facets of pain rather than trying to suppress them (adapted from https://commons.wikimedia.org/wiki/File:Kanizsa_triangle.svg; Accessed on March 14, 2018).

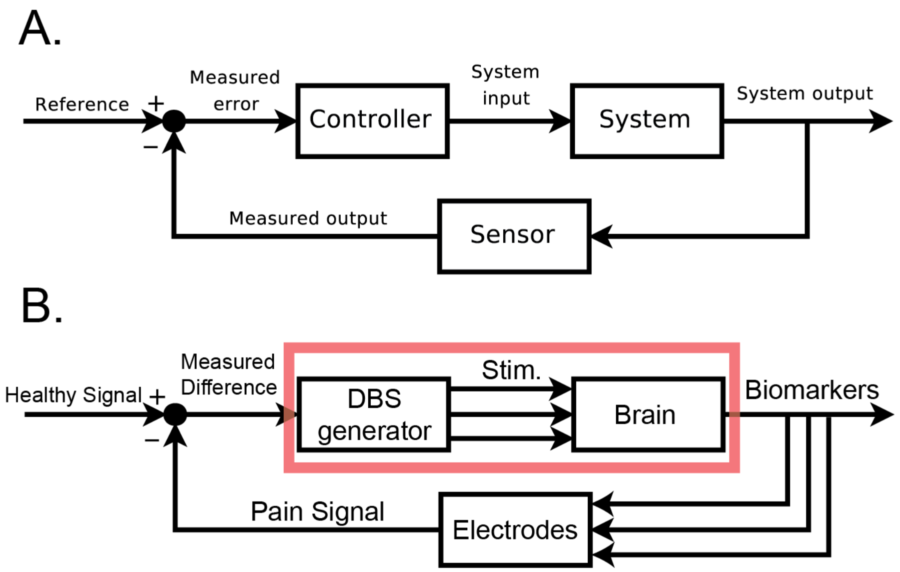


Figure 3.2. Block Diagram schematics of closed-loop control systems. (A) Classical block diagram of a single-input, single-output negative feedback control system, where the measured output of the system is compared to a reference signal via a closed-loop, to modify the system output and minimize error (adapted from [102]). (B) Example block diagram of a multi-input, multi-output closed-loop DBS system where a pain signal derived from biomarkers is compared to a reference signal via a feedback loop. Multi-regional stimulation is triggered to bring the system closer into the reference state. The red box highlights elements of an open-loop paradigm. aAvailable online at: https://upload.wikimedia.org/wikipedia/commons/2/24/Feedback_loop_with_descriptions.svg (Accessed Nov 30, 2017).

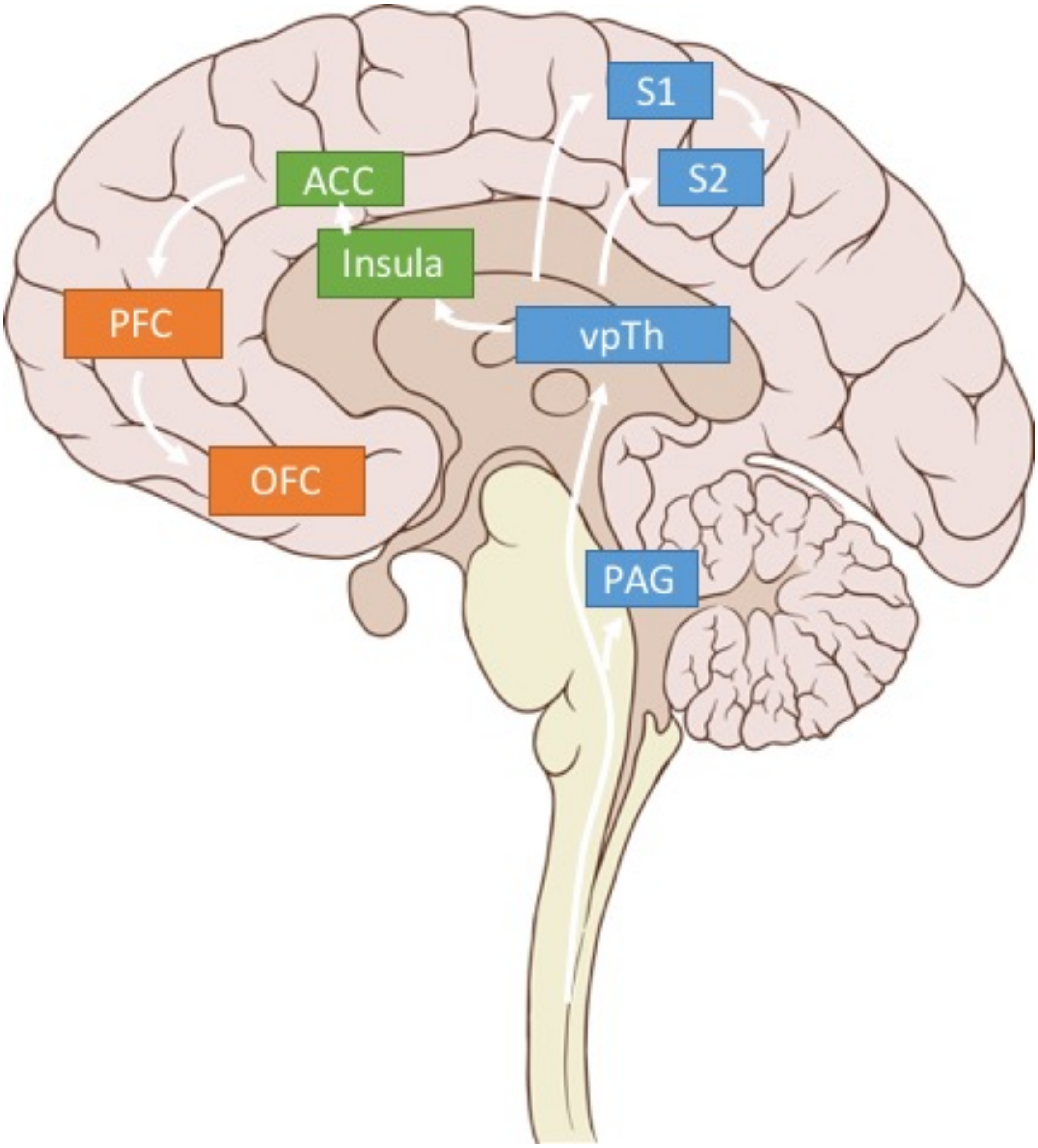


Figure 3.3. Pain related brain regions
(A) Key brain regions related to somatosensory (blue), affective (green) and cognitive (orange) pain processing. (Only regions of interest have been included for clarity).

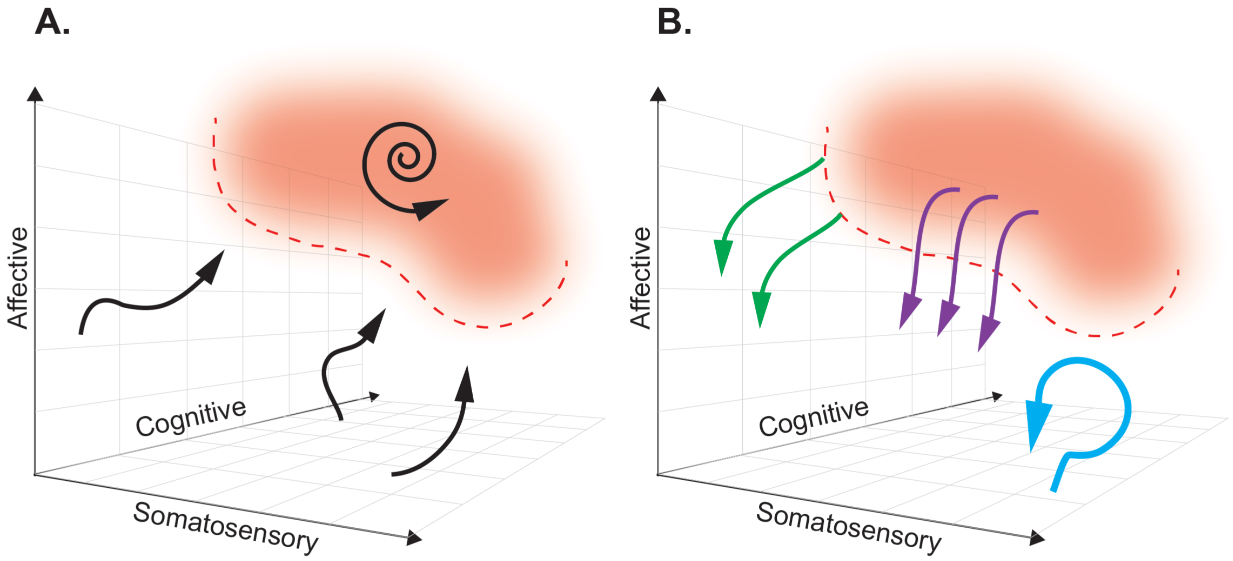


Figure 3.4. A multidimensional state space framework can be used to characterize pain states, reference states, and goals of DBS paradigms.
(A) A state space representing neural activity can be defined along the multiple dimensions of pain: somatosensory, affective and cognitive. For simplicity, a pain state is represented as a single red zone in the upper right corner, with defined threshold boundaries (dashed red line). The reference (pain-free) state is any region outside the red zone. The dynamics of neural activity that underlie transition from a pain-free state towards a pain state are shown as neural trajectories (black arrows). During constant baseline pain, there is a self-sustaining neural trajectory confined to the pain state (spiral arrow). (B) Different paradigms of DBS accomplish different goals. Tonic, open-loop DBS aims to maintain neural activity in a constant pain-free state (blue arrow). Abortive, patient-triggered or sensor-triggered DBS aims to push neural representations out of the pain state into the reference state (purple arrows). Closed-loop DBS will ideally deflect neural activity well before entering a pain-state (green arrows).

References

1. CDC. *Wide-ranging online data for epidemiologic research (WONDER)*. **2016**, (2016).

2. IASP. Classification of Chronic Pain, Second Edition (Revised) - IASP. Available at: https://www.iasp-pain.org/PublicationsNews/Content.aspx?ItemNumber=1673&navItemNumber=677. (Accessed: 5th December 2017)

3. Melzack, R. From the gate to the neuromatrix. *PAIN* **82**, S121 (1999).

4. Melzack, R. & Casey, K. Sensory, Motivational and Central Control Determinants of Pain: A New Conceptual Model. in *The Skin Senses* (1968).

5. Ossipov, M. H., Dussor, G. O. & Porreca, F. Central modulation of pain. *J. Clin. Invest.* **120**, 3779–3787 (2010).

6. Villemure, C. & Bushnell, C. M. Cognitive modulation of pain: how do attention and emotion influence pain processing? *Pain* **95**, 195–199 (2002).

7. Zubieta, J.-K. *et al.* Regional Mu Opioid Receptor Regulation of Sensory and Affective Dimensions of Pain. *Science* **293**, 311–315 (2001).

8. Dejerine, J. & Roussy, G. Le syndrome thalamique. *Rev Neurol Paris* **14**, 521–532 (1906).

9. Tasker, R. R. Thalamotomy. *Neurosurg. Clin. N. Am.* **1**, 841–864 (1990).

10. Wycis, H. T. & Spiegel, E. A. Thalamotomy and mesencephalothalamotomy; neuro-surgical aspects, including treatment of pain. *N. Y. State J. Med.* **49**, 2275–2277 (1949).

11. Shealy, C. N. Dorsal Column Electrohypalgesia. *Headache J. Head Face Pain* **9**, 99–102 (1969).

12. Adams, J. E., Hosobuchi, Y. & Fields, H. L. Stimulation of internal capsule for relief of chronic pain. *J. Neurosurg.* **41**, 740–744 (1974).

13. Hosobuchi, Y., Adams, J. E. & Rutkin, B. Chronic Thalamic Stimulation for the Control of Facial Anesthesia Dolorosa. *Arch. Neurol.* **29**, 158–161 (1973).

14. Adams, John E., Hosobuchi, Yoshio & Fields, Howard L. Stimulation of internal capsule for relief of chronic pain | Journal of Neurosurgery, Vol 41, No 6. (1974). Available at: http://thejns.org/doi/pdf/10.3171/jns.1974.41.6.0740. (Accessed: 2nd May 2017)

15. Levy, R., Deer, T. R. & Henderson, J. Intracranial neurostimulation for pain control: a review. *Pain Physician* **13**, 157–165 (2010).

16. Coffey, R. J. Deep brain stimulation for chronic pain: results of two multicenter trials and a structured review. *Pain Med. Malden Mass* **2**, 183–192 (2001).

17. Hosomi, K., Seymour, B. & Saitoh, Y. Modulating the pain network—neurostimulation for central poststroke pain. *Nat. Rev. Neurol.* **11**, 290–299 (2015).

18. Louppe, J.-M. *et al.* Motor cortex stimulation in refractory pelvic and perineal pain: Report of two successful cases. *Neurourol. Urodyn.* **32**, 53–57 (2013).

19. Brown, J. A. & Pilitsis, J. G. Motor cortex stimulation for central and neuropathic facial pain: a prospective study of 10 patients and observations of enhanced sensory and motor function during stimulation. *Neurosurgery* **56**, 290–297; discussion 290-297 (2005).

20. Lefaucheur, J.-P. *et al.* Motor cortex stimulation for the treatment of refractory peripheral neuropathic pain. *Brain J. Neurol.* **132**, 1463–1471 (2009).

21. Nauta, W. J. Hippocampal projections and related neural pathways to the midbrain in the cat. *Brain J. Neurol.* **81**, 319–340 (1958).

22. Papez, James. A proposed mechanism of emotion. *Arch. Neurol. Psychiatry* **38**, 725–743 (1937).

23. Ballantine, H. T., Cassidy, W. L., Flanagan, N. B. & Marino, R. Stereotaxic Anterior Cingulotomy for Neuropsychiatric Illness and Intractable Pain. *J. Neurosurg.* **26**, 488–495 (1967).

24. Whitty, C. W. M., Duffield, J. E., Tow, P. M. & Cairns, H. ANTERIOR CINGULECTOMY IN THE TREATMENT OF MENTAL DISEASE. *The Lancet* **259**, 475–481 (1952).

25. Foltz, E. L. & White, L. E. Pain “Relief” by Frontal Cingulumotomy. *J. Neurosurg.* **19**, 89–100 (1962).

26. Coghill, R. C., McHaffie, J. G. & Yen, Y.-F. Neural correlates of interindividual differences in the subjective experience of pain. *Proc. Natl. Acad. Sci.* **100**, 8538–8542 (2003).

27. Wager, T. D. *et al.* An fMRI-Based Neurologic Signature of Physical Pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).

28. Rainville, P., Duncan, G. H., Price, D. D., Carrier, B. & Bushnell, M. C. Pain Affect Encoded in Human Anterior Cingulate But Not Somatosensory Cortex. *Science* **277**, 968–971 (1997).

29. Johansen, J. P., Fields, H. L. & Manning, B. H. The affective component of pain in rodents: Direct evidence for a contribution of the anterior cingulate cortex. *Proc. Natl. Acad. Sci.* **98**, 8077–8082 (2001).

30. Johansen, J. P. & Fields, H. L. Glutamatergic activation of anterior cingulate cortex produces an aversive teaching signal. *Nat. Neurosci.* **7**, 398–403 (2004).

31. Spooner, J., Yu, H., Kao, C., Sillay, K. & Konrad, P. Neuromodulation of the cingulum for neuropathic pain after spinal cord injury. *J. Neurosurg.* **107**, 169–172 (2007).

32. Parvizi, J., Rangarajan, V., Shirer, W. R., Desai, N. & Greicius, M. D. The Will to Persevere Induced by Electrical Stimulation of the Human Cingulate Gyrus. *Neuron* **80**, 1359–1367 (2013).

33. Boccard, S. G. J., Pereira, E. A. C. & Aziz, T. Z. Deep brain stimulation for chronic pain. *J. Clin. Neurosci.* **22**, 1537–1543 (2015).

34. Boccard, S. G. J. *et al.* Long-Term Results of Deep Brain Stimulation of the Anterior Cingulate Cortex for Neuropathic Pain. *World Neurosurg.* **106**, 625–637 (2017).

35. Lempka, S. F. *et al.* Randomized clinical trial of deep brain stimulation for poststroke pain. *Ann. Neurol.* **81**, 653–663 (2017).

36. Bushnell, M. C., Čeko, M. & Low, L. A. Cognitive and emotional control of pain and its disruption in chronic pain. *Nat. Rev. Neurosci.* **14**, 502 (2013).

37. Hsieh, H. L. & Shanechi, M. M. Multiscale brain-machine interface decoders. in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 6361–6364 (2016). doi:10.1109/EMBC.2016.7592183

38. Smith, A. C. & Brown, E. N. Estimating a State-Space Model from Point Process Observations. *Neural Comput.* **15**, 965–991 (2003).

39. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).

40. Churchland, M. M., Yu, B. M., Ryu, S. I., Santhanam, G. & Shenoy, K. V. Neural Variability in Premotor Cortex Provides a Signature of Motor Preparation. *J. Neurosci.* **26**, 3697–3712 (2006).

41. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).

42. Flint, R. D., Lindberg, E. W., Jordan, L. R., Miller, L. E. & Slutzky, M. W. Accurate decoding of reaching movements from field potentials in the absence of spikes. *J. Neural Eng.* **9**, 046006 (2012).

43. Orsborn, A. L., So, K., Dangi, S. & Carmena, J. M. Comparison of neural activity during closed-loop control of spike- or LFP-based brain-machine interfaces. in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)* 1017–1020 (2013). doi:10.1109/NER.2013.6696109

44. So, K., Dangi, S., Orsborn, A. L., Gastpar, M. C. & Carmena, J. M. Subject-specific modulation of local field potential spectral power during brain–machine interface control in primates. *J. Neural Eng.* **11**, 026002 (2014).

45. Stavisky, S. D., Kao, J. C., Nuyujukian, P., Ryu, S. I. & Shenoy, K. V. A high performing brain–machine interface driven by low-frequency local field potentials alone and together with spikes. *J. Neural Eng.* **12**, 036009 (2015).

46. Chen, Z., Zhang, Q., Tong, A. P. S., Manders, T. R. & Wang, J. Deciphering neuronal population codes for acute thermal pain. *J. Neural Eng.* **14**, 036023 (2017).

47. Pandarinath, C. *et al.* High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife* **6**, e18554 (2017).

48. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012).

49. Schulz, E., Zherdin, A., Tiemann, L., Plant, C. & Ploner, M. Decoding an Individual’s Sensitivity to Pain from the Multivariate Analysis of EEG Data. *Cereb. Cortex* **22**, 1118–1123 (2012).

50. Kuo, P.-C., Chen, Y.-T., Chen, Y.-S. & Chen, L.-F. Decoding the perception of endogenous pain from resting-state MEG. *NeuroImage* **144**, 1–11 (2017).

51. Bokil, H., Andrews, P., Kulkarni, J. E., Mehta, S. & Mitra, P. Chronux: A Platform for Analyzing Neural Signals. *J. Neurosci. Methods* **192**, 146–151 (2010).

52. Hemptinne, C. de *et al.* Therapeutic deep brain stimulation reduces cortical phase-amplitude coupling in Parkinson's disease. *Nat. Neurosci.* **18**, 779 (2015).

53. Lieberman, M. D. & Eisenberger, N. I. The dorsal anterior cingulate cortex is selective for pain: Results from large-scale reverse inference. *Proc. Natl. Acad. Sci.* **112**, 15250–15255 (2015).

54. Reddan, M. C. & Wager, T. D. Modeling Pain Using fMRI: From Regions to Biomarkers. *Neurosci. Bull.* 1–8 (2017). doi:10.1007/s12264-017-0150-1

55. Goense, J., Bohraus, Y. & Logothetis, N. K. fMRI at High Spatial Resolution: Implications for BOLD-Models. *Front. Comput. Neurosci.* **10**, (2016).

56. Huster, R. J., Debener, S., Eichele, T. & Herrmann, C. S. Methods for Simultaneous EEG-fMRI: An Introductory Review. *J. Neurosci.* **32**, 6053–6060 (2012).

57. Apkarian, A. V., Bushnell, M. C., Treede, R.-D. & Zubieta, J.-K. Human brain mechanisms of pain perception and regulation in health and disease. *Eur. J. Pain* **9**, 463–463 (2005).

58. Gross, J., Schnitzler, A., Timmermann, L. & Ploner, M. Gamma Oscillations in Human Primary Somatosensory Cortex Reflect Pain Perception. *PLOS Biol.* **5**, e133 (2007).

59. Zhang, Z. G., Hu, L., Hung, Y. S., Mouraux, A. & Iannetti, G. D. Gamma-Band Oscillations in the Primary Somatosensory Cortex—A Direct and Obligatory Correlate of Subjective Pain Intensity. *J. Neurosci.* **32**, 7429–7438 (2012).

60. Sarnthein, J., Stern, J., Aufenberg, C., Rousson, V. & Jeanmonod, D. Increased EEG power and slowed dominant frequency in patients with neurogenic pain. *Brain* **129**, 55–64 (2006).

61. Stern, J., Jeanmonod, D. & Sarnthein, J. Persistent EEG overactivation in the cortical pain matrix of neurogenic pain patients. *NeuroImage* **31**, 721–731 (2006).

62. Boord, P. *et al.* Electroencephalographic slowing and reduced reactivity in neuropathic pain following spinal cord injury. *Spinal Cord* **46**, 118–123 (2008).

63. Schmidt, S. *et al.* Pain Ratings, Psychological Functioning and Quantitative EEG in a Controlled Study of Chronic Back Pain Patients. *PLOS ONE* **7**, e31138 (2012).

64. Ploner, M., Gross, J., Timmermann, L., Pollok, B. & Schnitzler, A. Pain Suppresses Spontaneous Brain Rhythms. *Cereb. Cortex* **16**, 537–540 (2006).

65. Nevian, T. The cingulate cortex: divided in pain. *Nat. Neurosci.* **20**, 1515 (2017).

66. Ploner, M., Sorg, C. & Gross, J. Brain Rhythms of Pain. *Trends Cogn. Sci.* **21**, 100–110 (2017).

67. Li, L. *et al.* Changes of gamma-band oscillatory activity to tonic muscle pain. *Neurosci. Lett.* **627**, 126–131 (2016).

68. Schulz, E. *et al.* Prefrontal Gamma Oscillations Encode Tonic Pain in Humans. *Cereb. Cortex* **25**, 4407–4414 (2015).

69. Wang, J.-Y., Luo, F., Chang, J.-Y., Woodward, D. J. & Han, J.-S. Parallel pain processing in freely moving rats revealed by distributed neuron recording. *Brain Res.* **992**, 263–271 (2003).

70. Moont, R., Crispel, Y., Lev, R., Pud, D. & Yarnitsky, D. Temporal changes in cortical activation during distraction from pain: A comparative LORETA study with conditioned pain modulation. *Brain Res.* **1435**, 105–117 (2012).

71. Liu, C.-C., Ohara, S., Franaszczuk, P. J., Crone, N. E. & Lenz, F. A. Attention to painful cutaneous laser stimuli evokes directed functional interactions between human sensory and modulatory pain-related cortical areas: *Pain* **152**, 2781–2791 (2011).

72. Ohara, S., Crone, N. E., Weiss, N. & Lenz, F. A. Analysis of synchrony demonstrates ‘pain networks’ defined by rapidly switching, task-specific, functional connectivity between pain-related cortical structures. *PAIN* **123**, 244 (2006).

73. Li, L. *et al.* Placebo Analgesia Changes Alpha Oscillations Induced by Tonic Muscle Pain: EEG Frequency Analysis Including Data during Pain Evaluation. *Front. Comput. Neurosci.* **10**, (2016).

74. Babiloni, C. *et al.* Anticipatory electroencephalography alpha rhythm predicts subjective perception of pain intensity. *J. Pain Off. J. Am. Pain Soc.* **7**, 709–717 (2006).

75. Tu, Y. *et al.* Alpha and gamma oscillation amplitudes synergistically predict the perception of forthcoming nociceptive stimuli. *Hum. Brain Mapp.* **37**, 501–514 (2016).

76. Kucyi, A. & Davis, K. D. The dynamic pain connectome. *Trends Neurosci.* **38**, 86–95 (2015).

77. Colgin, L. L. *et al.* Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature* **462**, 353–357 (2009).

78. Shirvalkar, P. R., Rapp, P. R. & Shapiro, M. L. Bidirectional changes to hippocampal theta–gamma comodulation predict memory for recent spatial episodes. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7054–7059 (2010).

79. Tort, A. B. L., Komorowski, R., Eichenbaum, H. & Kopell, N. Measuring Phase-Amplitude Coupling Between Neuronal Oscillations of Different Frequencies. *J. Neurophysiol.* **104**, 1195–1210 (2010).

80. Santaniello, S., Fiengo, G., Glielmo, L. & Grill, W. M. Closed-Loop Control of Deep Brain Stimulation: A Simulation Study. *IEEE Trans. Neural Syst. Rehabil. Eng.* **19**, 15–24 (2011).

81. Swann, N. C. *et al.* Gamma Oscillations in the Hyperkinetic State Detected with Chronic Human Brain Recordings in Parkinson's Disease. *J. Neurosci.* **36**, 6445–6458 (2016).

82. Keifer, O. P., Riley, J. P. & Boulis, N. M. Deep Brain Stimulation for Chronic Pain. *Neurosurg. Clin. N. Am.* **25**, 671–692 (2014).

83. Lima, M. C. & Fregni, F. Motor cortex stimulation for chronic pain: systematic review and meta-analysis of the literature. *Neurology* **70**, 2329–2337 (2008).

84. Ben-Menachem, E. *et al.* Vagus Nerve Stimulation for Treatment of Partial Seizures: 1. A Controlled Study of Effect on Seizures. *Epilepsia* **35**, 616–626 (1994).

85. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

86. Ashwin, P., Coombes, S. & Nicks, R. Mathematical Frameworks for Oscillatory Network Dynamics in Neuroscience. *J. Math. Neurosci.* **6**, (2016).

87. Rabinovich, M. I., Afraimovich, V. S., Bick, C. & Varona, P. Information flow dynamics in the brain. *Phys. Life Rev.* **9**, 51–73 (2012).

88. Radons, G., Becker, J. D., Dülfer, B. & Krüger, J. Analysis, classification, and coding of multielectrode spike trains with hidden Markov models. *Biol. Cybern.* **71**, 359–373 (1994).

89. Kumar, R. Methods for programming and patient management with deep brain stimulation of the globus pallidus for the treatment of advanced Parkinson's disease and dystonia. *Mov. Disord.* **17**, S198–S207 (2002).

90. Volkmann, J., Herzog, J., Kopper, F. & Deuschl, G. Introduction to the programming of deep brain stimulators. *Mov. Disord.* **17**, S181–S187 (2002).

91. Yang, Y. & Shanechi, M. M. Generalized binary noise stimulation enables time-efficient identification of input-output brain network dynamics. in *2016 38th*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 1766–1769 (2016). doi:10.1109/EMBC.2016.7591059

92. Hochberg, L. R. *et al.* Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* **442**, 164 (2006).

93. Sun, F. T. & Morrell, M. J. The RNS System: responsive cortical stimulation for the treatment of refractory partial epilepsy. *Expert Rev Med Devices* **11**, 563–72 (2014).

94. Swann, N. C. *et al.* Chronic multisite brain recordings from a totally implantable bidirectional neural interface: experience in 5 patients with Parkinson's disease. *J. Neurosurg.* 1–12 (2017). doi:10.3171/2016.11.JNS161162

95. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer-Verlag, 2009).

96. Ezzyat, Y. *et al.* Direct Brain Stimulation Modulates Encoding States and Memory Performance in Humans. *Curr. Biol.* **27**, 1251–1258 (2017).

97. Karamintziou, S. D. *et al.* Algorithmic design of a noise-resistant and efficient closed-loop deep brain stimulation system: A computational approach. *PloS One* **12**, e0171458 (2017).

98. Xiao, Y., Peña, E. & Johnson, M. D. Theoretical Optimization of Stimulation Strategies for a Directionally Segmented Deep Brain Stimulation Electrode Array. *IEEE Trans. Biomed. Eng.* **63**, 359–371 (2016).

99. Brocker, D. T. *et al.* Optimized temporal pattern of brain stimulation designed by computational evolution. *Sci. Transl. Med.* **9**, eaah3532 (2017).

100. Jazayeri, M. & Afraz, A. Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017).

101. Schultz, D. M. *et al.* Sensor-driven position-adaptive spinal cord stimulation for chronic pain. *Pain Physician* **15**, 1–12 (2012).

102. Orzetto. Control Theory. Available at: https://upload.wikimedia.org/wikipedia/commons/2/24/Feedback_loop_with_descriptions.svg. (Accessed: 30th November 2017)

Chapter 4: Discussion

Abstract

This dissertation was dedicated to understanding cross-area cortical communication in the context of learning. Specifically, it (1) examined cross-area communication in motor cortices during natural movement and BMI learning, and (2) hypothesized that pathological cross-area communication can lead to disorders such as chronic pain, and that disrupting that coordination provides potential for treatment.

This work is significant for its experimental methodological advances, scientific conclusions, and proposed hypotheses. Methodologically, it used a novel strategy to understand activity coordinated across two areas in a functional network; it demonstrated a viable approach to analyzing single-trial neural activity during variable learning behavior; and it directly addressed and analyzed off-target effects during local neural interventions. Scientifically, some important contributions of this work are updates to our understanding of the effect of learning on cross-area communication. Specifically, this work led to surprising conclusions about the differential impact of learning on the correlation versus the task-relevance of cross-area activity; introduced analyses of coordination between cross-area and local dynamics; and demonstrated the causality of local interventions on task encoding in distant brain regions. Finally, this work proposed a pragmatic and tractable translational framework for understanding and treating diseases that arise from pathological cross-area coordination, notably chronic pain.

Of course, there are still many limitations to this work, including the use of static dimensionality reduction methods which do not integrate important temporal information about neural activity; the narrow range of brain regions recorded, the limited number of both natural and BMI tasks explored, and the inherent constraints of proposing invasive neuromodulation as a therapeutic solution given the current functional limits of DBS devices approved for human use. Future work may remedy these limitations and expand results in new directions enabled by recent advances in recording and intervention technologies, new task paradigms for complex behaviors, and computational approaches designed to extract temporal dynamics in population activity.

This chapter expands on each of the above topics, namely (1) the significance of this dissertation, and (2) the limitations to the work and consequent future directions.

Significance

Methodological advances

Use of CCA to understand cross-area neural activity

Chapter 1 introduced the use of Canonical Correlation Analysis (CCA)[1] to extract and analyze neural dynamics shared between two brain regions, premotor (M2) and primary motor (M1) cortex, during motor skill learning. This approach departs from the traditional approach to understanding coordination of dynamics between brain regions, in which local, single-area signals are first reduced using dimensionality reduction techniques such as PCA or FA, and then compared [2–4]. While the traditional approach assumes that communicated signals are derived from locally aggregated activity, the approach we used instead leaves room for cross-area signals to be distinctly different from locally-dominant dynamics. The use of a similar method, Reduced Rank Regression (RRR) was previously used to analyze neural activity from V1 and V2 in anesthetized non-human primates and V1 and V4 in awake, behaving non-human primates [5]. However, in that study neural signals were first stripped of 'task-activity' by subtracting out trial-averaged stimulus-dependent signals. That strategy cannot be used to study neural signals during motor learning, as behavior is very variable, making trial-averaging particularly inappropriate. Additionally, as neural signals in the brain are neither trial-averaged or trial-mean subtracted, use of the raw neural signals seemed most biologically relevant to understanding mechanisms of plasticity during learning.

Analysis of single-trial neural signals during learning

Chapter 1 demonstrated an approach to understanding single-trial neural activity during variable learning behavior. Most studies relating neural activity to motor behavior rely on trial-averaging to produce de-noised versions of the neural activity [2,6]. To begin understanding behavioral variability some motor studies first group trials into tiers based on behavior (eg: slow, medium, fast movements) and then trial-average [7]. The use of modeling as a method for approximating de-noised single-trial neural signals has allowed for more fine-grained analysis of single-trial variable behavior [8–13]. Here we establish the tractability of combining dimensionality reduction methods with single-trial analysis to understand a dataset in which the overall variability of behavior is large.

Monitoring of off-target effects during local neural interventions

Chapters 1 and 2 use different strategies to directly inactivate and engage activity in a single node of the motor network while monitoring off-target effects in a partner region. In Chapter 1, we use muscimol to inactivate M2 and monitor the off-target effects on M1 activity and skilled reaching behavior. By recording activity before and during the intervention (i.e. baseline and inactivation sessions), we are able to first build models of cross-area activity during the baseline sessions, and then use those models to understand the effect of M2 inactivation on M1 neural activity. This is important because some studies without off-target monitoring have interpreted interventions as self-contained to the regions being inactivated [14]. However, other studies demonstrating off-target remodeling after chronic lesions as well as off-target disruptions during acute inactivation caution against the idea of purely 'local' interventions [15,16]. This more nuanced interpretation of local interventions also applies to the deliberate engagement of a specific brain region during tasks. In Chapter 2, we use BMI to train animals to modulate neural signals contained to M1. Several studies have examined how BMI learning affects both the neurons directly controlling the effector (i.e. direct units) as well as the other local neurons (i.e. indirect units) [17–19]. Here we additionally demonstrate that neurons in a distant brain region are engaged in task learning despite purportedly non-mandatory functional connectivity with local direct units.

Scientific advances

Learning differentially impacts correlation versus task-relevance of cross-area activity

Prior studies of learning-driven changes in cross-area communication have shown that learning increases the correlation of activity between partner regions. However, this work was mostly based on bulk local signals such as Local Field Potentials [20–23] (LFPs), in which fluctuations may be driven more by local dynamics than specific cross-area dynamics. Consequently, in those analyses, increases in coordinated task-modulation may confound (a) changes in cross-area coordination with (b) changes in local task-related modulation. In Chapter 1 we explicitly attempt to separate analyses quantifying the correlation of cross-area activity from analyses of the task-modulation of cross-area activity. We find that cross-area correlation does not change with learning, while task-modulation of cross-area activity increases. These findings suggest that correlation of activity across regions is not significantly driven by learning state, perhaps explaining animal's abilities to learn new movements quickly. Instead, learning may increase neural signaling corresponding to task-related movements within pre-existing connections already subserving other functions [24].

Learning increases coordination between local and cross-area dynamics

Our analyses of populations of neurons rather than bulk activity allowed us to distinguish local versus cross-area subspaces within each region's population activity. This approach had already been applied in a study of communication between visual cortex regions [5], which also found distinct local and cross-area subspaces (there referred to it as the 'communication subspace'). In that study, the authors hypothesized that changing the angle between the local and cross-area subspaces within the high-dimensional population neural space may drive cross-area coordination during learning. We directly tested this idea and found that the angle between local and cross-area subspaces did not change with learning. Instead, we found that task-modulation of local and cross-area dynamics became more similar with learning. This argues against the idea that communication between regions relies on assimilation of raw neural activity in those regions. Instead, communication between regions relies on coordinating the magnitude of relevant signals while still allowing the signals themselves to be *distinct*. The cross-area signals, as we identified them, serve as 'middlemen' for cross-area signaling by creating intermediate transforms between two distinct local signals.

The two above findings are reminiscent of Hebb's theories on adult learning [24]. In his seminal 1949 text, "The Organization of Behavior", Hebb distinguishes infant learning, in which many new connections are formed (i.e. neurons that fire together wire together), from adult learning, in which restructuring takes place within preexisting neural circuits. Specifically, he predicts that "The prompt learning of maturity is not an establishing of new connections but a selective reinforcement of connections already capable of functioning". In his model (Fig. 4.1), adult learning does not dramatically change inter-area connectivity, but instead restructures patterns of activity between local systems (A and B) and a 'transmission' subsystem (C). Until now, there has been little evidence supporting this theory. Even in adults, most studies report increased functional connectivity between regions with learning [20–23,25], and to our knowledge, no studies have attempted to distinguish 'subsystems' within ensemble representations during learning.

Local interventions affect task-encoding in distributed motor regions

Systems neuroscience concerns itself with the interactions of components within the nervous system. Despite this, studies using local inactivation [14] and chronic lesions have sometimes interpreted consequent behavior changes as resulting from circumscribed local damage and been unable to address the possibility of larger systemic dysfunction [26]. However, many studies of recovery after chronic lesions have shown extensive cortical remodeling [27,28], and recent studies have even compared these effects to acute inactivation [15]. Unfortunately, monitoring off-target effects of local interventions is rarely done [16]. Here we monitored off-target changes driven by two interventions: M2 local muscimol inactivation (Chapter 1) and M1-BMI learning (Chapter 2). We loosely consider BMI learning a 'local intervention' because we explicitly trained subjects to modulate neurons in a M1, without requiring changes in activity in M2. We found that M2 inactivation suppressed and disrupted M1 encoding of skilled reaching without affecting M1 mean firing or local shared variance; and M1-BMI learning drove task-modulation of neural activity in M2. These findings provide rare evidence that local interventions affect task-encoding in connected cortical regions.

Translational advances

The methods and findings included in this dissertation provide a framework for translational work geared at diagnosing neurological disorders and developing neuromodulatory therapies. From Chapter 1, similar analyses of cross-area cortical activity using CCA could improve diagnosis and management of disorders in patients with neural communication dysfunctions. For example, in Chapter 3 we outline how chronic pain can be framed as stemming from pathological coupling between sensory, affective, and cognitive regions engaged during pain, leading to inappropriate concordant activation of all three regions when any of them is engaged. To help distinguish chronic pain due to multi-area coupling with normal sensation from pain due to heightened responses to noxious stimuli, physicians could monitor activity in sensory, affective, and cognitive regions during pain experiences. Abnormal baseline coupling between regions might indicate the multi-area etiology, while abnormally elevated somatosensory activity without abnormal baseline multi-area coupling might indicate a more specific sensory etiology. If patients are diagnosed as having pathological multi-area coupling, a possible therapy would be to pair non-invasive stimulation of one region with non-invasive suppression of another region to promote de-coupling. A similar effect might be obtained through training by creating neurofeedback tasks designed to encourage patients to decrease or control cross-area coordination.

These approaches could potentially be applied to other neurological and psychiatric disorders hypothesized to affect specific functional networks. For example, one study of motor network function after unilateral subcortical strokes found that ipsilateral motor cortex became more interconnected with other nodes, while ipsilateral cerebellum became less connected [29]. Importantly, these measures were correlated with clinical measures of behavioral deficits. According to our model, paired stimulation of ipsilateral cerebellum with other nodes might aid in restoring function. Similarly, one study of brain-wide functional connectivity in patients with schizophrenia found global decreases in connectivity strength, along with increased diversity of functional connections [30]. Again, these measures were correlated with behavioral metrics. Based on those findings and the framework proposed in Chapter 3 for designing interventions, clinicians and neural engineers may be able to design new interventions using closed-loop multi-site neuromodulation.

Limitations and future directions

While the work in this dissertation made significant advances to common experimental and analytic methodologies, models of cross-area communication during learning, and frameworks for therapeutic neuromodulation, there are still many limitations to be addressed in future studies.

Limitations and future directions based on recording technologies

Chapters 1 and 2 use tungsten microwire probes to record simultaneous neural activity in M2 and M1 of rats during motor and BMI learning. The probes are each made of a 4 x 8 grid of wires covering approximately 1mm x 1.2 mm of cortex, allowing for relatively large spatial coverage of each brain region. However, there is an inherent tradeoff between using probes with large versus dense spatial coverage, such as those used in other recent studies of population recordings [31–33]. Due to the large spacing of wires, we do not obtain multiple views of neurons being recorded, limiting the ability to clearly sort single-neurons. Consequently, neuron activity or 'unit' activity referred to in this work often comes from several nearby units. However, this practice is common in studies of motor cortex in rodents [20,34–38], non-human primates [17,39–43], and humans[19,44]. Additionally, recent work in motor cortex of non-human primates has shown that latent signals in population activity can still be robustly extracted even from unsorted threshold crossings [45]. In addition to the limited spatial resolution of the neural activity, we did not track changes in single neurons over time, as is sometimes possible with optical recordings (though optical recordings have significantly lower temporal sampling) [3,7,25,32,46]. Future work would benefit from more dense recordings across a larger set of brain regions [32], with consistent tracking of neurons across learning. Of particular interest would be the evolution of cross-area activity between motor cortices and more cognitive regions, such as the prefrontal cortex; as well as with sub-cortical structures, such as the striatum[47]. Expanding the diversity of brain regions recorded would also allow for more extensive monitoring of off-target effects during interventions. Additionally, monitoring muscle activity during movements[17,48] may also contribute to understanding the role of cortical activity in movement execution, especially as several subcortical regions[47,49] and have been implicated in reach-to-grasp execution.

Limitations and future directions based on interventional methods

Chapter 1 used infusions of the GABA agonist muscimol to locally inactivate M2. This is an effective, widely-used method for dampening local activity[47,50]. However, the kinetics of muscimol spread lead to inactivation lasting at least several hours. Overall inactivation of M2 cell bodies also prevents isolation of neural signals transmitted specifically within M2 to M1 axon projections. More temporally and genetically precise activation and silencing of M2 cell bodies and/or terminals of cells that project to M1 may be possible with optogenetic methods[51]. Finally, Chapter 3 proposes a general framework for analyzing and treating chronic pain as a disorder of cross-area communication, but DBS, the main neuromodulatory method discussed, is invasive, expensive, and rarely accessible to patients. Alternatives strategies could include non-invasive stimulation, such as transcranial direct current stimulation (tDCS) or transcranial magnetic stimulation (TMS), both of which have been shown to have therapeutic effects in other systems disorders (eg: stroke rehabilitation[52] and depression[53], respectively). Focally targeted non-invasive neurostimulation is also actively being developed and have been shown to modulate local neural activity in rodent models [54]. However, non-invasive methods will need to be refined in order to be closed-loop, continuous, and portable in order to best treat intractable chronic disorders.

Limitations and future directions based on task paradigms

Chapters 1 and 2 analyzed neural data collected during learning of the reach-to-grasp task[55,56] (an example of skilled motor behavior), and a 1-dimensional BMI task. The reach-to-grasp behavior was chosen because it is kinematically similar to important grasping behaviors in humans and dependent on many cortical areas[57]. The BMI task[36], on the other hand, was designed be behaviorally simple and require learned modulation of very few, explicitly chosen neurons[58]. Examining neural activity during a range of tasks is important for the generalizability of our overarching interpretations. To continue testing generalizability, future work could examine cross-area interactions during types of learning which are instead dependent on sensory functional networks (eg: sensory discrimination tasks), cognitive networks (eg: memory-based decision-making tasks), affective networks (eg: fear conditioning), or a combination therein (eg: set-shifting).

While use of animal models allows for greater availability of subjects and flexibility of experimental paradigms, it is also important to compare findings from human subjects who can provide introspective reports of their learning and behavioral strategies. For example, a recent study of BMI learning in a human patient asked the subject to try different cognitive strategies to control a cursor (eg: imagining moving her wrist towards or away from the target location). By comparing the patient's neural activity with her verbal reports of her strategy, the study was able to corroborate the theory that local covariance structures in motor cortex (i.e. intrinsic variables) correspond to explicit and accessible cognitive strategies for movement control [19]. Similarly, future work on cross-area coordination during learning in humans may demonstrate that the "feeling" of being triggered to perform certain skilled movements in specific contexts (e.g. getting "in the zone" to swim when you get to the pool) depends on widespread coordinated cross-area activity in sensory, cognitive, affective, and motor regions.

Limitations and future directions based on computational methods

Our results are bound by the assumptions and limitations in our analysis methods and our approaches for modeling task-relevant activity. In Chapter 2, task-related activity is modeled as the trial-averaged mean of activity aligned to the end of the trials, which reflects modulation used to perform the BMI task. In contrast, in Chapter 1, task-related activity is modeled using static dimensionality reduction methods which extract relationships in the variance shared between either local and cross-area populations of neurons (using FA and CCA, respectively). While trial-averaging emphasizes the important of activity relative to specific moment in time, FA and CCA emphasize population covariances in a manner agnostic to timing.

To be able to jointly examine temporal and covariance relationships in populations, it is necessary to use methods which take into account temporal relationships of activity within a population (i.e. sequences, or 'tiling' of information). Recent development of two methods, Gaussian-Process Factor Analysis (GPFA) [9], and seqNMF [59], offer such opportunities. GPFA can capture shared population variance even when neurons are not co-active (unlike FA), and it returns the timescales over which the activity is distributed. This information could in turn be used to optimize parameter selection in analyses using seqNMF, which captures repeated sequences in population activity, given an expectation for sequence duration. Together, analyses using these computational methods may allow for a better understanding of structured information across interacting cortical regions, such as M1 and M2. In particular, future work could examine whether co-activating, sequentially binding, or even interleaving task-based sequences across areas correlates with task learning.

Multi-area neural data from BMI tasks seem particularly appropriate for such analyses. Other approaches to understanding emergence of structured activity could include comparison of population dynamics within and across regions in early versus late learning within single days, during which it is more feasible to maintain the same set of neurons. Finally, analyses in this dissertation are centered on understanding experimentally collected neural data. However, modeling interactions between artificial populations [60–62] of interconnected neurons (i.e. in silico modeling) could help elucidate guiding principles of cross-area interactions during learning.

Understanding cross-area neural dynamics on a single-trial basis in experiments is critical to later designing closed-looped, multi-area neuromodulatory therapies which will rely on real-time evaluation of complex, multi-faceted brain states.

Figures

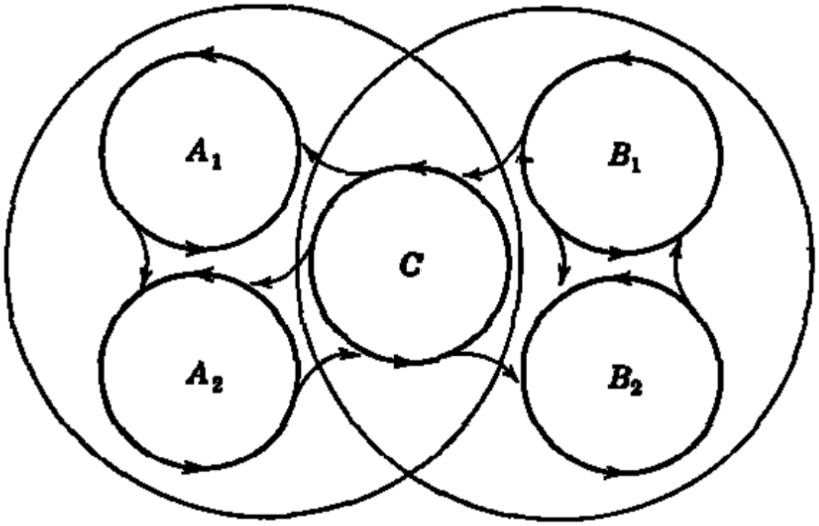


Figure 4.1. Hebb's model for two-system learning.
(Reproduced from Hebb, 1949) [24]. "To illustrate the possibility that a subsystem, C, may act as a link between two systems (conceptual complexes). One concept is represented by A_1, A_2, and C, the second by B_1, B_2, and C. The two systems have a subsystem, C, in common, to provide a basis of prompt association. "

References

1. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28**, 321–377 (1936).

2. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* **17**, 440 (2014).

3. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).

4. Perich, M. G., Gallego, J. A. & Miller, L. E. A Neural Population Mechanism for Rapid Learning. *Neuron* **100**, 964-976.e7 (2018).

5. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical Areas Interact through a Communication Subspace. *Neuron* (2019). doi:10.1016/j.neuron.2019.01.026

6. Chandrasekaran, C., Peixoto, D., Newsome, W. T. & Shenoy, K. V. Laminar differences in decision-related neural activity in dorsal premotor cortex. *Nat. Commun.* **8**, 614 (2017).

7. Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).

8. Kiani, R., Cueva, C. J., Reppas, J. B. & Newsome, W. T. Dynamics of Neural Population Responses in Prefrontal Cortex Indicate Changes of Mind on Single Trials. *Curr. Biol.* **24**, 1542–1547 (2014).

9. Yu, B. M. *et al.* Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *J. Neurophysiol.* **102**, 614–635 (2009).

10. Churchland, M. M., Yu, B. M., Sahani, M. & Shenoy, K. V. Techniques for extracting single-trial activity patterns from large-scale neural recordings. *Curr. Opin. Neurobiol.* **17**, 609–618 (2007).

11. Kao, J. C. *et al.* Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nat. Commun.* **6**, 7759 (2015).

12. Pandarinath, C. *et al. Latent factors and dynamics in motor cortex and their application to brain-machine interfaces*. (PeerJ Inc., 2018). doi:10.7287/peerj.preprints.27217v1

13. Wei, Z., Inagaki, H., Li, N., Svoboda, K. & Druckmann, S. An orderly single-trial organization of population dynamics in premotor cortex predicts behavioral variability. *Nat. Commun.* **10**, 216 (2019).

14. Brown, A. R. & Teskey, G. C. Motor Cortex Is Functionally Organized as a Set of Spatially Distinct Representations for Complex Movements. *J. Neurosci.* **34**, 13574–13585 (2014).

15. Otchy, T. M. *et al.* Acute off-target effects of neural circuit manipulations. *Nature* **advance online publication**, (2015).

16. Jazayeri, M. & Afraz, A. Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017).

17. Ganguly, K., Dimitrov, D. F., Wallis, J. D. & Carmena, J. M. Reversible large-scale modification of cortical networks during neuroprosthetic control. *Nat. Neurosci.* **14**, 662–667 (2011).

18. Gulati, T., Guo, L., Ramanathan, D. S., Bodepudi, A. & Ganguly, K. Neural reactivations during sleep determine network credit assignment. *Nat. Neurosci.* **advance online publication**, (2017).

19. Sakellaridi, S. *et al.* Intrinsic Variable Learning for Brain-Machine Interface Control by Human Anterior Intraparietal Cortex. *Neuron* **102**, 694-705.e3 (2019).

20. Koralek, A. C., Long, J. D., Costa, R. M. & Carmena, J. M. Corticostriatal dynamics during learning and performance of a neuroprosthetic task. in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2682–2685 (2010). doi:10.1109/IEMBS.2010.5626632

21. DeCoteau, W. E. *et al.* Learning-related coordination of striatal and hippocampal theta rhythms during acquisition of a procedural maze task. *Proc. Natl. Acad. Sci.* **104**, 5644–5649 (2007).

22. Benchenane, K. *et al.* Coherent Theta Oscillations and Reorganization of Spike Timing in the Hippocampal- Prefrontal Network upon Learning. *Neuron* **66**, 921–936 (2010).

23. Arce-McShane, F. I., Ross, C. F., Takahashi, K., Sessle, B. J. & Hatsopoulos, N. G. Primary motor and sensory cortical areas communicate via spatiotemporally coordinated networks at multiple frequencies. *Proc. Natl. Acad. Sci.* **113**, 5083–5088 (2016).

24. Hebb, D. The organization of behavior: A neuropsychological theory. (1949).

25. Makino, H. *et al.* Transformation of Cortex-wide Emergent Properties during Motor Learning. *Neuron* **94**, 880-890.e8 (2017).

26. Kawai, R. *et al.* Motor Cortex Is Required for Learning but Not for Executing a Motor Skill. *Neuron* **86**, 800–812 (2015).

27. Jones, T. A. & Adkins, D. L. Motor System Reorganization After Stroke: Stimulating and Training Toward Perfection. *Physiology* **30**, 358–370 (2015).

28. Ramanathan, D., Conner, J. M. & H. Tuszynski, M. A form of motor cortical plasticity that correlates with recovery of function after brain injury. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11370–11375 (2006).

29. Wang, L. *et al.* Dynamic functional reorganization of the motor execution network after stroke. *Brain J. Neurol.* **133**, 1224–1238 (2010).

30. Lynall, M.-E. *et al.* Functional Connectivity and Brain Networks in Schizophrenia. *J. Neurosci.* **30**, 9477–9487 (2010).

31. Chung, J. E. *et al.* High-Density, Long-Lasting, and Multi-region Electrophysiological Recordings Using Polymer Electrode Arrays. *Neuron* **101**, 21-31.e5 (2019).

32. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. 13 (2019).

33. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).

34. Ramanathan, D. S., Gulati, T. & Ganguly, K. Sleep-Dependent Reactivation of Ensembles in Motor Cortex Promotes Skill Consolidation. *PLoS Biol* **13**, e1002263 (2015).

35. Ramanathan, D. S. *et al.* Low-frequency cortical activity is a neuromodulatory target that tracks recovery after stroke. *Nat. Med.* **24**, 1257–1267 (2018).

36. Gulati, T., Ramanathan, D. S., Wong, C. C. & Ganguly, K. Reactivation of emergent task-related ensembles during slow-wave sleep after neuroprosthetic learning. *Nat. Neurosci.* **17**, 1107–1113 (2014).

37. Hyland, B. Neural activity related to reaching and grasping in rostral and caudal regions of rat motor cortex. *Behav. Brain Res.* **94**, 255–269 (1998).

38. Kargo, W. J. Improvements in the Signal-to-Noise Ratio of Motor Cortex Cells Distinguish Early versus Late Phases of Motor Skill Learning. *J. Neurosci.* **24**, 5560–5569 (2004).

39. Lara, A. H., Cunningham, J. P. & Churchland, M. M. Different population dynamics in the supplementary motor area and motor cortex during reaching. *Nat. Commun.* **9**, 2754 (2018).

40. Churchland, M. M. *et al.* Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010).

41. Hatsopoulos, N. G., Xu, Q. & Amit, Y. Encoding of Movement Fragments in the Motor Cortex. *J. Neurosci.* **27**, 5105–5114 (2007).

42. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

43. Golub, M. D. *et al.* Learning by neural reassociation. *Nat. Neurosci.* (2018). doi:10.1038/s41593-018-0095-3

44. Pandarinath, C. *et al.* High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife* **6**, e18554 (2017).

45. Trautmann, E. M. *et al.* Accurate estimation of neural population dynamics without spike sorting. *bioRxiv* 229252 (2017). doi:10.1101/229252

46. Cao, V. Y. *et al.* Motor Learning Consolidates Arc-Expressing Neuronal Ensembles in Secondary Motor Cortex. *Neuron* (2015). doi:10.1016/j.neuron.2015.05.022

47. Lemke, S. M., Ramanathan, D. S., Guo, L., Won, S. J. & Ganguly, K. Emergent modular neural control drives coordinated motor actions. *Nat. Neurosci.* 1 (2019). doi:10.1038/s41593-019-0407-2

48. Hyland, B. I. & Jordan, V. M. B. Muscle activity during forelimb reaching movements in rats. *Behav. Brain Res.* **85**, 175–186 (1997).

49. Esposito, M. S., Capelli, P. & Arber, S. Brainstem nucleus MdV mediates skilled forelimb motor tasks. *Nature* **508**, 351–356 (2014).

50. Tian, L. Y. & Brainard, M. S. Discrete Circuits Support Generalized versus Context-Specific Vocal Learning in the Songbird. *Neuron* **96**, 1168-1177.e5 (2017).

51. Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M. & Deisseroth, K. Optogenetics in Neural Systems. *Neuron* **71**, 9–34 (2011).

52. Kang, N., Weingart, A. & Cauraugh, J. H. Transcranial direct current stimulation and suppression of contralesional primary motor cortex post-stroke: a systematic review and meta-analysis. *Brain Inj.* **32**, 1063–1070 (2018).

53. Chen, J. *et al.* Left versus right repetitive transcranial magnetic stimulation in treating major depression: A meta-analysis of randomised controlled trials. *Psychiatry Res.* **210**, 1260–1264 (2013).

54. Grossman, N. *et al.* Noninvasive Deep Brain Stimulation via Temporally Interfering Electric Fields. *Cell* **169**, 1029-1041.e16 (2017).

55. Whishaw, I. Q. & Pellis, S. M. The structure of skilled forelimb reaching in the rat: A proximally driven movement with a single distal rotatory component. *Behav. Brain Res.* **41**, 49–59 (1990).

56. Wong, C. C., Ramanathan, D. S., Gulati, T., Won, S. J. & Ganguly, K. An automated behavioral box to assess forelimb function in rats. *J. Neurosci. Methods* **246**, 30–37 (2015).

57. Whishaw, I. Q., Pellis, S. M., Gorny, B. P. & Pellis, V. C. The impairments in reaching and the movements of compensation in rats with motor cortex lesions: an endpoint, videorecording, and movement notation analysis. *Behav. Brain Res.* **42**, 77–91 (1991).

58. Fetz, E. E. Operant conditioning of cortical unit activity. *Science* **163**, 955–958 (1969).

59. Mackevicius, E. L. *et al.* Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *eLife* **8**, e38471 (2019).

60. Goudar, V. & Buonomano, D. V. Encoding sensory and motor patterns as time-invariant trajectories in recurrent neural networks. 28

61. Hardy, N. F. & Buonomano, D. V. Neurocomputational models of interval and pattern timing. *Curr. Opin. Behav. Sci.* **8**, 250–257 (2016).

62. Williamson, R. C. *et al.* Scaling Properties of Dimensionality Reduction for Neural Populations and Network Models. *PLOS Comput. Biol.* **12**, e1005141 (2016).

Appendices

Appendix A: Do Ask, Do Tell: UCSF SOM asks applicants about sexual orientation and gender identity

Authors: B. Perkovich[1,†] and T.L. Veuthey[1, 2, 3, †]

[1]School of Medicine, University of California San Francisco, San Francisco CA, USA.

[2]Neuroscience Graduate Program, University of California San Francisco, San Francisco CA.

[3]Medical Scientist Training Program, University of California San Francisco, San Francisco CA.

[†] These authors contributed equally to this work.

Article Type: Op-Ed

This year, for the first time, the UCSF School of Medicine is including on its secondary application optional questions about sexual orientation and gender identity. The change was the result of a multi-year collaboration between the School of Medicine's Office of Admissions and students from the Lesbian, Gay, Bisexual, Transgender, and Queer Student Alliance (LGBTQSA). It arose from a shared desire to make the admissions experience more welcoming to students who identify with the LGBTQ community. Allowing applicants to self-identify as LGBTQ enables the school to collect critical data and to send an important message about UCSF's commitment to an inclusive campus climate.

Within its mandatory primary application, the American Medical College Application Service (AMCAS), routinely collects information on many aspects of an applicant's background, including sex (as male, female, or decline to state), age, race, national and ethnic origin, citizenship, place of legal residence, and socioeconomic status. However, information on sexual orientation and non-traditional gender identity is not captured. UCSF School of Medicine is now presenting an opportunity for applicants to voluntarily provide that information on its secondary application. Data on LGBTQ identity will allow UCSF to better understand the needs of its applicant population. Currently, this information is restricted to the Office of Admissions; however, if students could opt to make their information available to the Office of Diversity, targeted outreach efforts could be developed. LGBTQ-identified applicants could be connected with current student mentors, and incoming students could be introduced to the wealth of resources available on campus to support the LGBTQ community. Ultimately, such programs would better enable UCSF to attract the most talented and diverse students, including those who identify as LGBTQ.

Applications are an opportunity for students to present how their experiences can contribute to a campus climate. LGBTQ-identified people should feel safe in discussing their backgrounds on admissions applications, including how identity informs their perspectives and professional ambitions. However, for many LGBTQ people, experiences with stigma and discrimination have taught them not to discuss their identity openly. For others, LGBTQ-identified or not, sexual orientation and gender identity may not seem relevant to their application. Applicants will choose for themselves what information they will provide. No matter how an individual responds, the application is now more reflective of the diversity and inclusiveness at the core of the UCSF experience.

In 2007, the University of California Board of Regents released a Diversity Statement that highlighted “the acute need to remove barriers to the recruitment, retention, and advancement of talented students, faculty, and staff from historically excluded populations.” LGBTQ-identified people have been, are, and will continue to be an important part of the fabric of UCSF. However, there was previously no systematic way of collecting information about sexual orientation and gender identity as incoming students entered into our community. With these new questions, the School of Medicine is better able to understand its applicants and is better able to communicate that applicants’ identities and past experiences matter here. By asking optional questions inclusive of LGBTQ identity, the UCSF School of Medicine can communicate that differences in sexual orientation and gender identity are a part of the diversity that makes UCSF such a vibrant place.

The new questions on sexual orientation and gender identity, like all questions on the UCSF secondary application, are works in progress. If you are interested in contributing to the discussion around these questions, or are interested in getting more involved in LGBTQ student life at UCSF, get in touch with the LGBTQSA and consider attending the opening meeting of the organization this coming fall.

Appendix B: Tiptoeing around it: Inference from absence in potentially offensive speech

Authors: M. A. Gates[1], T. L. Veuthey[2], M.H. Tessler[3], K. A. Smith[4], T. Gerstenberg[4], L. Bayet[5], J. B. Tenenbaum[4]

Affiliations:

[1] Psychology, University of California, Berkeley

[2] Neuroscience, University of California, San Francisco

[3] Psychology, Stanford University

[4] Brain and Cognitive Sciences, Massachusetts Institute of Technology

[5] Laboratories of Cognitive Neuroscience, Harvard Medical School & Boston Children's Hospital

Article Type: Original Research

Keywords: politeness; social meaning; pragmatics; Bayesian cognitive model; Rational Speech Act model

Abstract

Language that describes people in a concise manner may conflict with social norms (e.g., referring to people by their race), presenting a conflict between transferring information efficiently and avoiding offensive language. When a speaker is describing others, we propose that listeners consider the speaker's use or absence of *potentially offensive language* to reason about the speaker's goals. We formalize this hypothesis in a probabilistic model of polite pragmatic language understanding, and use it to generate predictions about interpretations of utterances in ambiguous contexts, which we test empirically. We find that participants are sensitive to potentially offensive language when resolving ambiguity in reference. These results support the idea that listeners represent conflicts in speakers' goals and use that uncertainty to interpret otherwise underspecified utterances.

Introduction

Referring to strangers can be challenging. Without knowing their name, you could describe them by their physical appearance, but not all attributes are equally informative. One problem for speakers is that highly diagnostic attributes can be potentially offensive (e.g., an overweight person's weight).

Grice (1975, p. 46)[1] was aware of this problem: "There are, of course, all sorts of other maxims (aesthetic, social, or moral in character), such as 'Be polite', that are also normally observed by participants in talk exchanges." In a politeness framework, the avoidance of potentially offensive words illustrates how speakers balance being informative with social goals [2]. Specifically, Brown and Levinson (1987) outline ambiguous speech as a form of indirect or "off-record" politeness. We draw inspiration from these ideas and hypothesize that the use or avoidance of words that carry social meaning prompts listeners to reason about the speaker's social goals. Do listeners hypothesize that speakers are constrained to use inoffensive language, and use this understanding to infer a speaker's intended meaning from an ambiguous utterance?

We developed a model in the Rational Speech Act (RSA) tradition [3,4] to capture the social and epistemic inferences elicited by words with social meaning, specifically potentially offensive descriptors. Vanilla RSA models predict pragmatic inferences listeners make for literally ambiguous statements by considering the alternative statements the speaker could have said. Recent work has modeled inferences about speakers' *social* goals, specifically the desire to be kind to the listener (Polite RSA[5,6]). The polite RSA model defines the social utility of an utterance as the quality of the world it makes the listener believe they are in. We extend this work by having potentially offensive utterances incur a *social cost* to the speaker. A listener who is aware of these social costs can resolve otherwise ambiguous utterances to infer a speaker's intended referent.

In our experiments, participants were introduced to a world where the words "blue" or "green" were potentially offensive. With their new social understanding, they played reference games in which they were asked to interpret a speaker's utterance (e.g., "person with the hat") in terms of which character in a scene the speaker was trying to refer to (see Figure A.1).

We hypothesize that listeners reason about the social cost of producing potentially offensive speech a) to contextually understand ambiguous utterances, and b) to evaluate speakers. Experiment 1 tests participants' inferences about who an ambiguous utterance refers to. Experiment 2 measured participants' inferences about the speaker's goals. Across these two experiments, we find that our model accounts for the fine-grained inferences listeners draw when reasoning about potentially offensive speech.

Computational Model

We built a rational model of communication within the Rational Speech Act framework [3,4]. Our model belongs to the class of "uncertain RSA" models, which involve reasoning about aspects of the speaker beyond just their intended meaning[7]. We used this framework to understand the phenomenon wherein a speaker is underinformative so as to not use potentially offensive speech, but listeners are nevertheless able to infer who speakers are referring to. In other words, when listeners are aware of a speaker's alternative utterances and the associated social costs, they can reason backwards to infer the speaker's intended referent.

Specifically, this work builds on an RSA model for polite language use (Polite RSA[6]). The listener in Polite RSA reasons about whether the speaker was trying to be epistemically informative (à la Vanilla RSA) or considerate to the listener's feelings (a social goal). The Polite RSA model operationalizes the social utility of an utterance *u* in terms of the subjective value of the world state that the listener would believe themselves to be in upon hearing *u*. For example, positive social utility is incurred by making the listener believe they are in a good state (e.g., that the cookies they baked were delicious). The model predicts that speakers who try to balance being informative and kind will choose to produce more *indirect speech* (e.g., saying "it wasn't amazing" as opposed to "it was terrible"), and this prediction was borne out empirically[6].

We took inspiration from the Polite RSA model, but parametrized the reasoning slightly differently. We modeled a listener who reasons about a potential social cost to an utterance. That is, words could be costlier to produce by the speaker by virtue of their social stigma of use. We assumed, for example, that a socially-minded speaker would incur a cost by referring to an overweight person as "fat". Rather be on the word form itself, this kind of cost can likely be derived out of a more basic mechanism analogous to that used by Yoon et al. (2017)[6], a point we return to in the Discussion.

Model details

The RSA framework models utterances and inferences as deriving from recursive social reasoning: a speaker S_1 produces an utterance u reasoning about how a literal listener L_0 would interpret it. A pragmatic listener L_1 interprets the utterance u reasoning about what speaker S_1 would say.

We start with the literal listener L_0, who literally interprets the meaning of any utterance u to determine the intended referent r within the context C:

Equation (1): $P_{L_0}(r \mid u, C) \propto [[\, f(u)]]\,(r) \cdot P(r)$

$[[u]]\,(r)$ is u's literal meaning, mapping to 1 if u matches referent r and 0 otherwise given context C. $f(u)$ expresses the noisy semantics model: with probability γ the listener doesn't condition on the utterance heard and instead samples a referent from the prior[8,9]. Mathematically, $P\, f(u_{w-1})|\, f(u) = 1 - \gamma,\ \forall\ w \in u$, where each w represents a word in the utterance u. $P(r)$ is a uniform distribution over possible referents given the context C.

Speaker S_1 produces an utterance based on a utility function U, which has two parts. The first part represents an epistemic utility which we define as the literal listener L_0

uncertainty about the referent *r* after hearing the utterance *u*: ln $P_{L_0}(r \mid u, C)$. This uncertainty is weighted by an utterance prior *P*(*u*) that assigns more probability to utterances with fewer words (uttering words is effortful). If $\sum_W$ is the utterance's word count, *W* is the maximum number of words possible in an utterance, and ξ parameterizes, then $P(u)$ = $^{[\exp(-\xi\cdot\sum w)]]}$ /[$\sum_{w=0:W}$ $\exp(-\xi\cdot\sum w)$]. We introduce a weighting parameter β_{epi} which captures how much the speaker cares about reducing the listener's uncertainty about the true referent.

The second part of speaker S_1's utility function represents a social utility. In our experiments and model, color terms are potentially offensive. The speaker is aware of a specific color word which is considered potentially offensive and designated as *badWord*. The speaker's social utility is $V(u) = 0$ if $badWord \in u$, and 1 otherwise. We introduce another weighting parameter β_{soc} which captures how much the speaker cares about avoiding potentially offensive language. By combining both epistemic and social utility, we get S_1's utility function as follows:

$U(u, r, C, \beta\,\hat{}\,\beta) = \beta_{epi} \cdot \ln P_{L_0}(r \mid u,C) \cdot P(u) + \beta_{soc} \cdot V(u)$

Overall, the speaker chooses an utterance softmax-optimally, where λ_1 represents S_1 's optimality:

Equation (2): $P_{S_1}(u|r, C, \beta) \propto \exp \lambda_1 \cdot U(u, r, C, \beta)$

The pragmatic listener L_1 then reasons about the speaker S_1, jointly inferring the referent *r* and how much weight the speaker S_1 places on the epistemic β_{epi} and social

β_{SOC} utility[10]. $P(r)$ is uniform over possible referents given context C, and $P(\hat{\beta})$ is a uniform distribution across the set {0.1, 0.3, 0.5, 0.7, 0.9}.

Equation (3): $P_{L_1}(r, \hat{\beta} \,|u, C) \propto P_{S_1}(u|r, C, \hat{\beta}) \cdot P(r) \cdot P(\hat{\beta})$

We implemented the model in WebPPL, a probabilistic programming language[11]. The model has three free parameters: a parameter for the noisy semantics (i.e., the overall extent to which utterances are not truth-functional) γ, a cost to producing more words ξ, and the speaker optimality parameter λ_1. In parameter fitting, γ was fixed at .1, and the other parameters were fit to the data, but restricted to the following ranges (consistent with models of the same model class): ξ fell between 0-1 and λ_1 fell between 1-20 [5,6]. The best-fitting parameter settings were: $\xi = 0.5$, and $\lambda_1 = 20$, determined through minimizing the least-squared error between model predictions and behavioral results.

In our experiments, utterances *u* could be any combination of the following: n/a (in the experiment, we added “person” to all utterances, so participants saw “the person” instead), one color term (“blue”, “green”, or “orange”), “scarf”, and “hat”. So, for example, an utterance could be “the person” or “the orange person with the scarf”. The intended referent *r* could be any of the two or three possible referents that appeared within a context *C*. The potentially offensive color term *badWord* was either “blue” or “green”, counterbalanced across participants.

We tested our model against human behavior in two experiments. In Expt. 1, listeners inferred the intended referent *r* given an utterance *u* and context *C*. In Expt. 2, listeners inferred $\hat{\beta}$ given a referent *r*, utterance *u*, and context *C*.

General Experiment Methods

Participants

We recruited participants from Amazon Mechanical Turk, with U.S. IP addresses and no reported color blindness. In Experiment 1, 45 participants were recruited, and three were removed (two for later reporting colorblindness, and one for failing a catch-trial). In Experiment 2, 46 participants were recruited, and one removed for later reporting colorblindness.

Stimuli and Procedure

Training: Participants began by viewing training scenes. Training scenes were designed to inform participants that using a particular color (*badWord*: either “blue” or “green”) was potentially offensive. Participants first read and were tested on an explicit description of the manipulation: “In a parallel world, some people are different colors. In this world, calling someone a ‘[color] person’ is potentially offensive,” where [color] was *badWord*. Participants then viewed several counterbalanced scenes, in which characters were selectively scolded by other characters for saying *badWord*.

Main Experiment: Following the training scenes, participants viewed reference game contexts in the main experiment. Within each context, two or three people were aligned left to right, were colored blue, green, or orange, and possibly wore hats and scarves. In the accompanying text, participants were observing the possible referents with a speaker named [Name]. [Name]s were selected by random selection with replacement from a list of 172 names for each context. The order of the possible referents in the context was randomly sampled at the beginning of the experiment and was fixed for all participants. The order of trials was randomized.

Context selection: Contexts were selected to test the inference that if a speaker did not explicitly refer to a person by their color, then perhaps that color was a *badWord* and potentially offensive. We sought examples that produced a range of model predictions. Contexts were selected to be roughly consistent across the experiments, so that the different methods of probing potential offensiveness could be compared. Finally, contexts were chosen to have built-in controls, such that if an image was presented where the referent color was *badWord*, the same image type was presented in a different context where the colors were switched so that the referent was now not *badWord*. In the rest of this paper, we describe the analysis with respect to the *badWord* being “blue”.

Experiment 1: Inferring the referent

Experiment-Specific Methods

This experiment contained 35 contexts. In each context, a speaker presented an utterance and the participant was asked to select which of the 2-3 referents the speaker was likely referring to (simple multiple-choice task, see Figure A.1).

Results and Discussion

In calculating statistics, because probabilities for the last referent of each context were entirely determined by probabilities assigned to the other referent(s), values from a randomly chosen referent were removed from further statistical analyses in order to meet assumptions of independence.

Participants' social inferences closely mirrored the inferences predicted by the model. Specifically, if the speaker's statement was ambiguous, participants selected the person with the potentially offensive color as being the referent, as predicted by the model (e.g. see contexts 1A, 1G in Table 1). When no referents of potentially-offensive colors were available, participants and the model were approximately ambivalent between the referents (e.g. see context 1B). In "positive control" contexts, in which the referent was unambiguously indicated by an utterance describing the intended referent's color ("the blue person"), participants selected the designated referent, as predicted by the model (e.g. see context 1F).

The left plot in Figure A.2 shows a scatterplot with model predictions and participants' inferences across all contexts. Our model explained participants' inferences to a high degree of quantitative accuracy with bootstrapped 95% confidence intervals for adjusted R^2 of [0.86, 0.96] and for Spearman's ρ of [0.88, 0.97]. The model's incorporation of social utility was critical to fit participants' inferences: when social utility was removed in a lesioned version of the model, bootstrapped confidence intervals for adjusted R^2 dropped to [0.27, 0.60], and for Spearman's ρ to [0.62, 0.89] (Figure A.2, right). Moreover, the moderately high correlations from the lesioned model were mostly driven by the presence of the positive control contexts in Expt. 1, which did not require social knowledge. When the eight positive control contexts were removed, the lesioned model's bootstrapped confidence intervals for adjusted R^2 dropped to [0.06,0.43], and for Spearman's ρ to [0.37, 0.83].) 10000 samples were drawn in all cases.

While the model generally captured participants' inferences well, there was a subset of contexts for which the model's predictions did not match participants' inferences. In these contexts, the model was reluctant to make the inference that the speaker was referring to the person with the potentially-offensive color when that person wore an item which was not specified in the utterance. Specifically, first consider the normal case: in context 1D, the only way to pick out the blue person would be to refer to their color. Given this fact, upon hearing "the person" instead, the model correctly predicted that people would choose the blue person as the intended referent. However, in context 1C, the blue person could also be unambiguously identified by referring to their scarf. Upon hearing the utterance "the person" in this context, the model was unsure who the intended referent was, whereas people considered the blue person with the scarf to be most likely. A similar phenomenon occurred in context 1E. One possible explanation for the deviation between model predictions and people's judgments here is that participants may have learned to associate the utterance “the person” with a blue person based on inferences drawn in previous contexts.

Experiment 2: Inferring speaker goals

We placed participants in a world where certain words were potentially offensive in Expt. 1. Given this knowledge, we found that listeners could infer a speaker's intended referent even if the speaker was ambiguous, as predicted by the model. In Expt. 2, we tested whether listeners could infer a speaker's goals (informational or social) based on how the speaker referred to someone.

Experiment-Specific Methods

After viewing the same training scenes that participants had seen Experiment 1, participants saw additional training scenes that clarified that the dimension of "offensiveness" corresponded to the use of *badWord*, and that the dimension of "ambiguity" referred to how much the utterance specifically identified the intended referent. Participants answered a comprehension check question, and then saw 40 different contexts in the test phase. Figure A.3 shows a screenshot of the test phase. In each context, an intended referent (out of two or three possible referents) was circled, and two possible utterances the speaker could say were shown on the left and right sides of the screen. Participants moved two separate sliders ranging from 0 to 100 to indicate which of the two utterances they considered to be more offensive, and which to be more ambiguous. The sliders were initially set at 50, which represented ambivalence.

Results and Discussion

Similarities between model predictions and judgments: Overall, the model again provided an accurate account of participants' inferences. If one utterance better distinguished the referent, participants rated that utterance as less ambiguous, as predicted by the model. This rating of lower ambiguity appeared over relatively subtle distinctions, like when the utterance reduced the number of valid possible intended referents from 3 to 2 (see for example context 2G in Table 2) or from 2 to 1 (e.g. contexts 2D, 2E). If utterances were both equally informative, participants roughly rated them as equally informative (e.g. context 2B) though behavioral exceptions exist.

With respect to offensiveness, if a single utterance contained the word "blue", then that utterance was rated as more offensive (e.g. context 2F). If neither utterance contained "blue", those utterances were rated as equally (un)offensive (e.g. context 2G). If both utterances contained the word "blue", then those utterances were roughly rated as equally offensive (e.g. context 2D), but see minor trends below.

Overall, model predictions and participants' judgments were highly correlated (Figure A.4). Bootstrapped confidence intervals (alpha = .025, adjusted for multiple comparisons, 10^4 samples) for adjusted R^2 were [.72,.90] for ambiguity and [.90,.98] for offensiveness; for Spearman's ρ intervals were [.85,.96] for ambiguity and [.66,.90] for offensiveness.

Differences between model predictions and judgments: The behavioral responses did, however, differ from the model in a few systematic ways. An important trend that occurred in behavior was that people found utterances to be much less informative if redundant traits were not listed (e.g. context 2H). While the model predicted that saying “the person with the hat” and “the person” would be equally informative if all possible referents were wearing hats, participants found “the person” to be much more ambiguous. While this desire for “redundant overinformativity” is not captured in our model, it is often observed in referent games[8].

However, some preference for information redundancy was indeed captured by the model through the noisy semantics assumption. In context 2C, the model predicted that an utterance with two informative words is less ambiguous than an utterance with one informative word — because a listener with “noisy hearing” might miss one.

Another systematic divergence between model predictions and participants’ judgments was that when asked about ambiguity, the model engaged in social inference more than participants did (e.g. contexts 2A, 2F). For example, if an utterance was “the person” when one possible referent was blue and the other green, the model made the social inference that the speaker was trying to refer to the blue person and predicted the utterance “the person” to be less ambiguous than it would have been without the social inference. However, in this setup, participants rarely appeared to make this inference. Instead, participants seemed to treat the ambiguity question as separate from the knowledge they were demonstrating in the offensiveness question (in which they were indicating that the term “blue” was potentially offensive.)

The results comparing the full model to the lesioned model (social utility set to 0) support the above hypothesis. When the social considerations were removed, the model predictions for ambiguity became *closer* to the behavioral results (e.g. context

2A). Numerically, for ambiguity ratings, bootstrapped confidence intervals (alpha = .025, 10^4 samples) for adjusted R^2 were [0.78, 0.95] for the lesioned model (compared to [0.72, 0.90] for the full model) and for Spearman's ρ were [0.91, 0.98] for the lesioned model (compared to [0.85, 0.96] for the full model). (The equivalent comparison with the lesioned model for offensiveness ratings was trivial by design, as the lesioned model was always ambivalent over utterances.)

The finding that participants did not engage social reasoning when asked about ambiguity may be due to question framing. "Offensiveness" and "ambiguity" ratings were clearly delineated in Experiment 2, and the focus on answering each separately (in addition to the extra training scenes that differentiated them) may have discouraged social reasoning to crossover into inferences about ambiguity.

On the offensiveness question, the differences between model predictions and participants' judgments were relatively small. Interestingly, participants considered any mention of color as slightly more offensive than model predictions, even if that color was non-offensive (e.g. contexts 2B, 2C). Participants also considered it slightly more offensive to say a color term if no other features were mentioned (e.g. contexts 2D, 2E), or to say "the person" alone (e.g. context 2H). These results are intuitive: if a feature like "blue" is offensive, it suggests that the general category of color might be avoided; and it feels rude to not say anything when referencing someone. Future work will probe how to add these intuitions into a richer, hierarchical model that draws generalizations ("don't refer to color") from specific instances ("don't say blue").

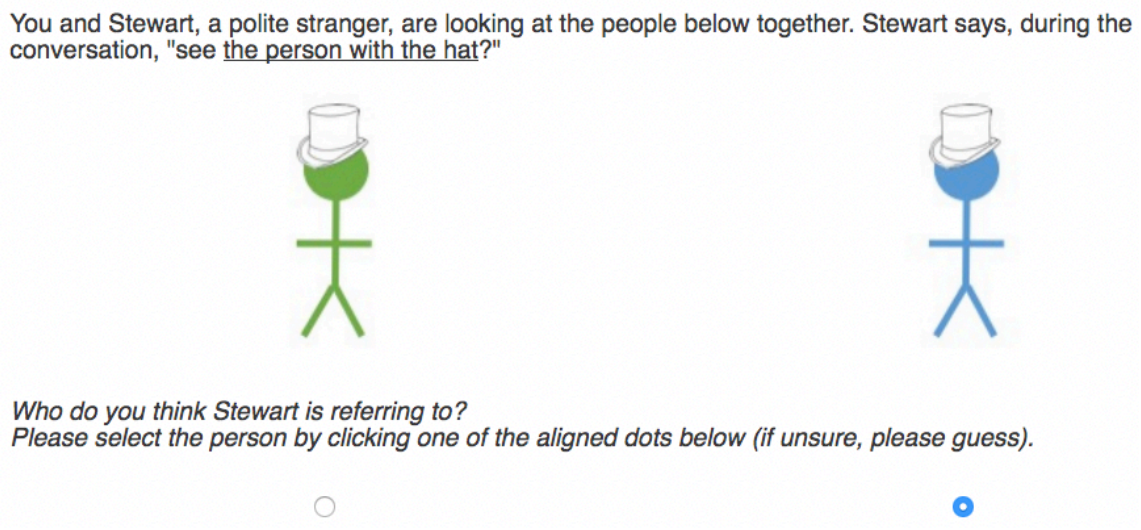
Conclusion

Some words are potentially offensive. This means that in some situations, the most efficient way of referring to someone may incur a social cost, creating a tension between efficiency and social adeptness of speech. We hypothesized that when listeners and speakers have shared knowledge of this tension, speakers can avoid using offensive speech and listeners can resolve otherwise ambiguous utterances to correctly infer the speaker's intended referent.

To make these ideas precise, we built on an existing model of polite language understanding by introducing a *social cost* that a speaker incurs for producing potentially offensive language. The model captures the inference that people make in determining a speaker's intended referent given an utterance that is ambiguous but constrained by social cost (Experiment 1), and also captures the explicit access that participants have to a speaker's epistemic and social goals given their utterance and context (Experiment 2). This work shows how the general mechanism of reasoning about the social function of language employed by the speaker[5,6] can begin to explain how listeners reason from the absence of potentially offensive language to resolve reference in context. While the model overall provides a very good fit to participants' inferences and judgments in both experiments, there were also some discrepancies which motivate future extensions of the model.

In our model, we directly mark potentially offensive words with a social utterance cost, but the same word might be offensive in one context and not another, or if said by one speaker but not another. One possibility is that it is a derivative property of subjective values associated with world states, in the style of Yoon et al. (2016)[5], perhaps by speakers putting themselves in the listener’s shoes and imagining themselves being referred to in a particular way. Another possibility is that these costs arise from social signaling: the speaker does not want the listener to infer that they are the type of person that calls people “blue”. In future work we hope to investigate how the social cost of potentially offensive speech is grounded in the complex social inferences that listeners and speakers draw about each other.

Figures



You and Stewart, a polite stranger, are looking at the people below together. Stewart says, during the conversation, "see the person with the hat?"

Who do you think Stewart is referring to?
Please select the person by clicking one of the aligned dots below (if unsure, please guess).

○ ●

Figure A.1. Example context from Experiment 1.

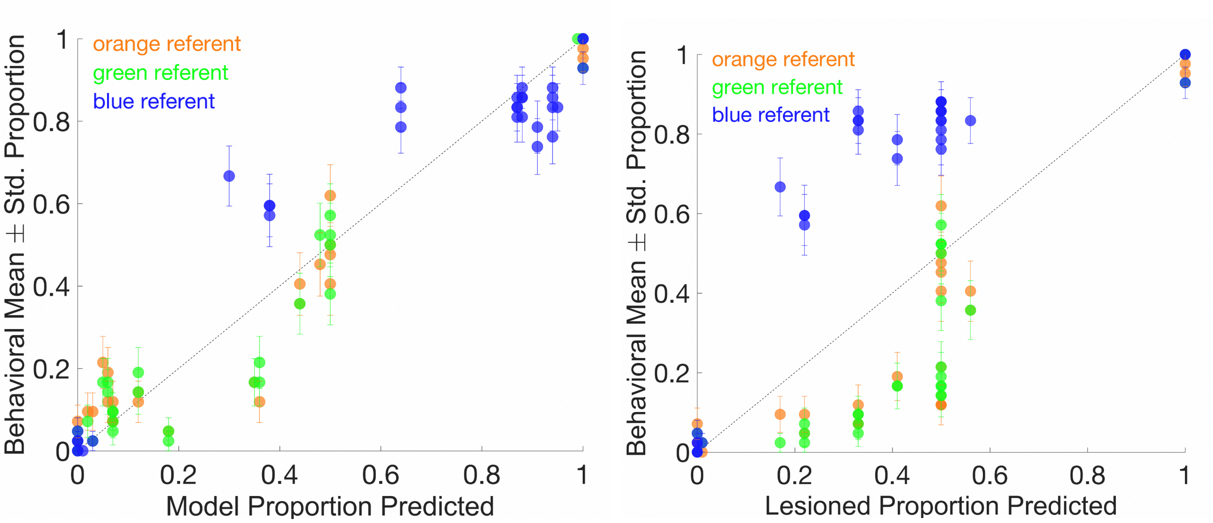
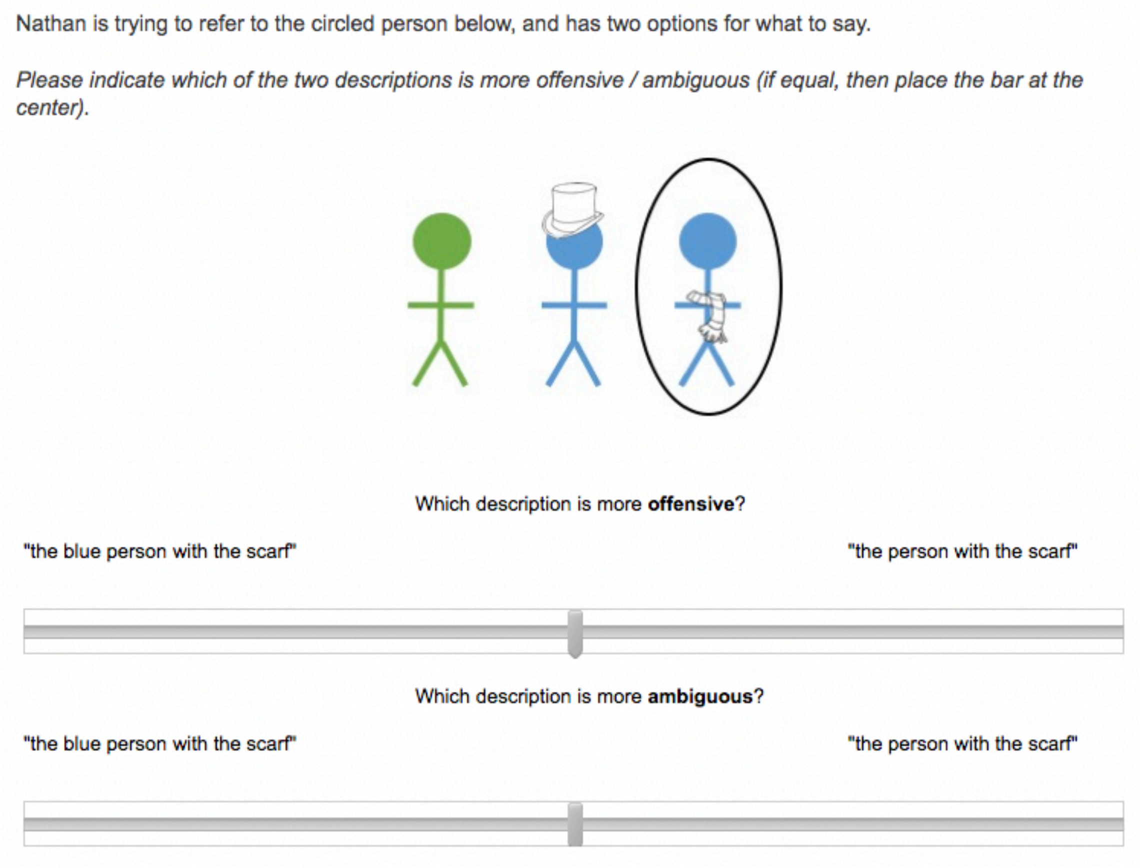


Figure A.2. Behavioral and model comparison for Expt. 1.
Participants saw an utterance and inferred which of the 2-3 referents the speaker was referring to for 35 contexts. Referents were orange, green, or blue. Results were collapsed across conditions so that "blue" was the potentially offensive word in all contexts. Behavioral results show the proportion of participants selecting each referent; model predictions show the proportions that the model allocated to each referent. Left: Full model. Right: Lesioned model (social utility set to 0).



Nathan is trying to refer to the circled person below, and has two options for what to say.

Please indicate which of the two descriptions is more offensive / ambiguous (if equal, then place the bar at the center).

Which description is more **offensive**?

"the blue person with the scarf" "the person with the scarf"

Which description is more **ambiguous**?

"the blue person with the scarf" "the person with the scarf"

Figure A.3: Example context from Experiment 2.

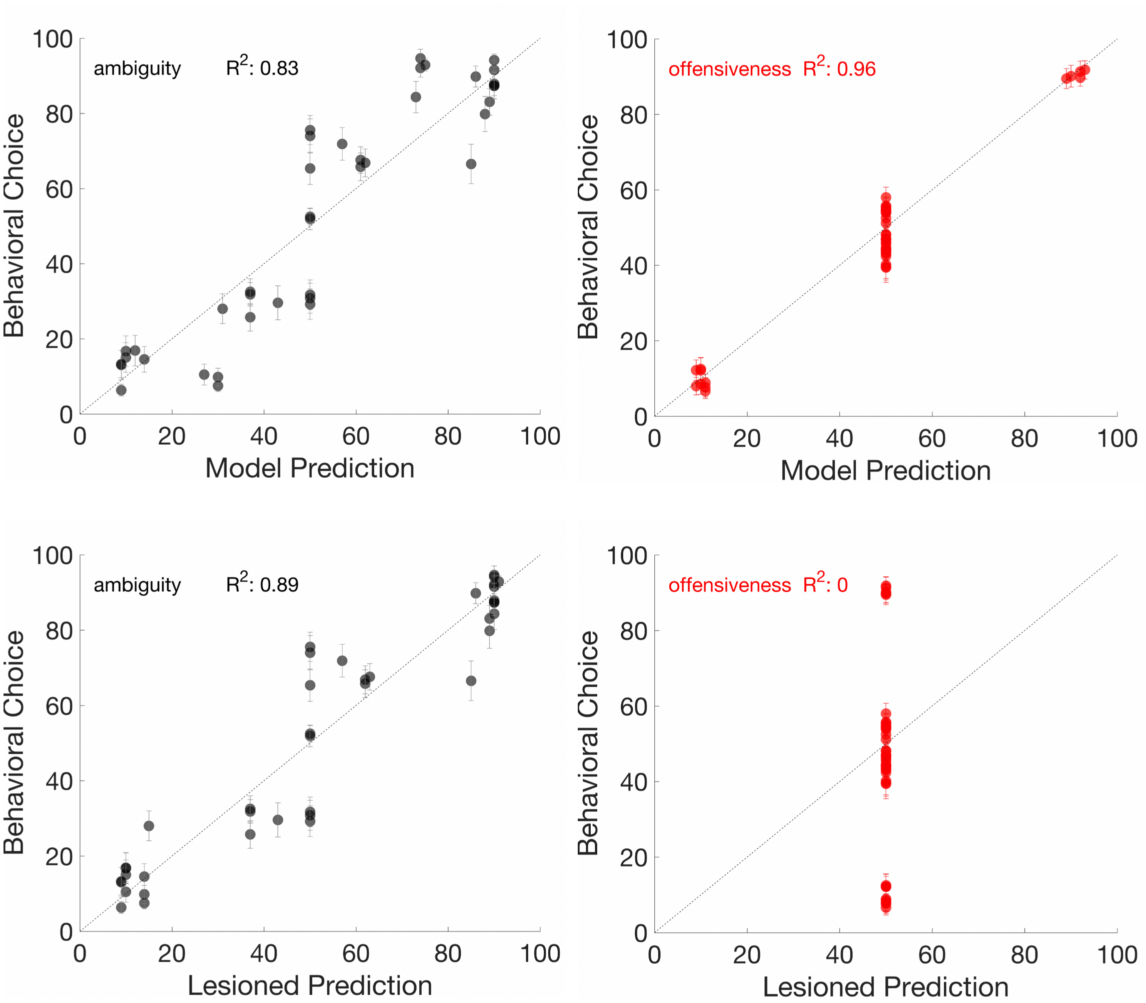


Figure A.4. Behavioral and model comparison for Experiment 2.
Participants rated which of two utterances describing a scene was more ambiguous (left), and which was more offensive (right) in 40 contexts. Behavioral results are the mean and standard error of participants' ratings of utterances, ranging from 0 (the utterance to the left of the screen was rated most ambiguous/offensive) to 100 (the utterance to the right was rated most ambiguous/offensive). Thus, lower scores indicate that the left utterance was rated more highly (more ambiguous / offensive) than the right utterance, and higher scores indicate that the right utterance was rated more highly than the left utterance. Model responses are the rescaled difference between β_{epis} / β_{soc} for the left and right utterances. Adjusted R^2 values are reported. Top: Full model. Bottom: Lesioned model (social utility set to 0).

Tables

	Referents (*Bl=potentially offensive)	Utterance	Behavioral Mean *(Std.)* Props.	Model Props.	Lesioned Props.
1A	Or / Bl	“n/a”	.14 *(.05)* / .86 *(.05)*	.12 / .88	.5 / .5
1B	Gr-hat / Or-scarf	“n/a”	.50 *(.08)* / .50 *(.08)*	.5 / .5	.5 / .5
1C	Or / Bl-scarf / Gr	“n/a”	.17 *(.06)* / .67 *(.07)* / .17 *(.06)*	.35 / .30 / .35	.41 / .17 / .41
1D	Or-scarf / Bl / Gr-hat	“n/a”	.10 *(.05)* / .83 *(.06)* / .07 *(.04)*	.02 / .95 / .02	.22 / .56 / .22
1E	Bl-hat / Or / Gr-scarf	“n/a”	.57 *(.08)* / .40 *(.08)* / .02 *(.02)*	.38 / .44 /.18	.22 / .56 / .22
1F	Bl-scarf-hat / Gr / Or-scarf-hat	“blue hat scarf”	.93 *(.04)* / 0 / .07 *(.04)*	1 / 0 / 0	1 / 0 / 0
1G	Bl-scarf-hat / Or-scarf-hat / Gr-scarf-hat	“hat scarf”	.81 *(.06)* / .12 *(.05)* / .07 *(.04)*	.87 / .07 /.07	.33 / .33 / .33

Table 1. Example Expt. 1 contexts.
For each context, the 2-3 referents are separated by “/” and can be blue (“Bl”), green (“Gr”), or orange (“Or”). Results were collapsed across conditions so that “blue” was the potentially offensive word in all contexts. In the “Utterance” column, “n/a” stands in for “the person”, and “blue hat scarf” for “the blue person with the hat and the scarf”. The behavioral, model, and lesioned model (without social inference) proportions allocated to each referent are shown.

	Referents (*Bl=potentially offensive)	Utt. 1	Utt. 2	Amb	AmbM	AmbL	Off	OffM	OffL
2A	Gr-scarf **/ Bl-scarf**	“blue scarf”	“scarf”	92 *(2)*	74	90	9 *(2)*	11	50
2B	Bl-scarf **/ Gr-hat**	“green”	“hat”	52 *(3)*	50	50	43 *(3)*	50	50
2C	**Gr-hat** / Bl-scarf	“hat”	“green hat”	33 *(3)*	37	37	55 *(3)*	50	50
2D	**Bl-scarf** / Bl-hat / Gr	“blue”	“blue scarf”	13 *(4)*	9	9	40 *(3)*	50	50
2E	**Gr-scarf** / Gr-hat / Bl	“scarf”	“green”	83 *(4)*	89	89	55 *(3)*	50	50
2F	Gr-hat **/ Bl-hat /** Bl-scarf-hat	“n/a”	“blue”	7 *(1)*	30	14	91 *(3)*	92	50
2G	**Gr-hat** / Gr-scarf-hat / Bl-hat	“hat”	“green hat”	15 *(3)*	14	14	52 *(3)*	50	50
2H	Bl-hat / Gr-scarf-hat **/ Gr-hat**	“hat”	“n/a”	74 *(5)*	50	50	54 *(2)*	50	50

Table 2. Example Experiment 2 contexts.
For each context, referents are separated by slashes (the intended referent is in bold) and could be blue (“Bl”), green (“Gr”), or orange (“Or”). Results were collapsed across conditions so that “blue” was the potentially offensive word in all contexts. Each context had two utterances: “Utt. 1” was positioned on the left of the screen at score 0, and “Utt. 2” was positioned on the right at score 100. Thus, lower scores indicate that Utt. 1 was rated higher (more ambiguous / offensive) than Utt. 2, and higher scores indicate Utt. 2 was rated higher (more ambiguous / offensive) than Utt. 1. In the experiment, these utterances were longer than the abbreviations shown here: “the person” was shown rather than “n/a”, and “the blue person with the scarf” rather than “blue scarf”. In the results columns, “Amb” indicates ambiguity ratings: behavioral mean and italicized standard errors are shown (“Amb”), as are model predictions (“AmbM”) and lesioned model predictions without social inference (“AmbL”). “Off” indicates offensiveness ratings.

Acknowledgments

This work was supported by NSF STC award CCF-1231216 to the Center for Brains, Minds & Machines, the Marine Biological Lab. in Woods Hole, the Berkeley AI Travel Stipend & NIH Berkeley Neuro. Grant to MAG, the NDSEG & the UCSF Discovery Fellowship to TV, a Philippe Foundation grant to LB, & NSF Graduate Research Fellowship DGE-114747 to MHT.

References

1. Grice, H. P. Logic and Conversation. *Semant.-Pragmat. Bound. Philos.* **47**, (1975).

2. Brown, P. & Levinson, S. C. Politeness: Some universals in language usage. *Camb. Uni Press* **4**, (1987).

3. Frank, M. C. & Goodman, N. D. Predicting Pragmatic Reasoning in Language Games. *Science* **336**, 998–998 (2012).

4. Goodman, N. D. & Stuhlmüller, A. Knowledge and implicature: Modeling language understanding as social cognition. *Top. Cogn. Sci.* **5**, 173–184 (2013).

5. Yoon, E. J., Tessler, M. H., Goodman, N. D. & Frank, M. C. Talking with tact: Polite language as a balance between kindness and informativity. in *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (Cognitive Science Society, 2016).

6. Yoon, E. J., Tessler, M. H., Goodman, N. D. & Frank, M. C. “I won’t lie, it wasn’t amazing”: Modeling polite indirect speech. in *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (Cognitive Science Society, 2017).

7. Goodman, N. D. & Frank, M. C. Pragmatic Language Interpretation as Probabilistic Inference. *Trends Cogn. Sci.* **20**, 818–829 (2016).

8. Degen, J., Hawkins, R. X. D., Graf, C., Kreiss, E. & Goodman, N. D. When redundancy is rational: A Bayesian approach to ‘overinformative’ referring expressions. *ArXiv190308237 Cs* (2019).

9. Graf, C., Degen, J., Hawkins, R. X. D. & Goodman, N. D. Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions. 6

10. Goodman, N. D. & Lassiter, D. Probabilistic Semantics and Pragmatics Uncertainty in Language and Thought. in *The Handbook of Contemporary Semantic Theory* (eds. Lappin, S. & Fox, C.) 655–686 (John Wiley & Sons, Ltd, 2015). doi:10.1002/9781118882139.ch21

11. Goodman, N. D. & Stuhlmüller, A. The Design and Implementation of Probabilistic Programming Languages. (2014). Available at: http://dippl.org/. (Accessed: 11th June 2019)

Appendix C: Why you need an agenda for meetings with your principal investigator

Authors: T.L. Veuthey[1, 2, †] and S. Thompson[3,†]

[1]Neuroscience Graduate Program, University of California San Francisco, San Francisco CA.

[2]Medical Scientist Training Program, University of California San Francisco, San Francisco CA.

[3]Biophysics Graduate Program, University of California San Francisco, San Francisco CA, USA.

[†] These authors contributed equally to this work.

Article Type: Op-Ed

As PhD students, we often find ourselves discussing our interactions with our principal investigators (PIs) and swapping advice for improving our mentoring meetings. We have found three practices to be consistently helpful: asking our PIs about all aspects of their job; preparing an agenda for each meeting; and negotiating new experiments without explicitly saying ‘no’.

We both see our PhD programmes as academic apprenticeships. One crucial goal is to flesh out our understanding of life as a PI. By collaborating with our PIs and observing how they work, we learn how to plan experiments and how to write papers. But we don’t get to practise other skills, such as interacting with journal editors and recruiting lab members. To learn these, we ask our PIs about how they plan when running the lab. For example, when people leave Samuel’s lab, he asks his PI about her plans for reallocating shared lab responsibilities.

Face-to-face time with our PIs must be focused, so we use agendas to organize the conversation. We habitually start with, “I made a list of topics I wanted to talk to you about.” Tess often starts her agendas with an update on her efforts to develop new research equipment so that her PI can evaluate their importance to her project. When Tess was designing new probes for electrophysiological recordings, her PI helped her to balance testing new research hardware against continuing data collection with older technology. Preparing an agenda also helps us to learn our PIs’ priorities. Before Samuel discusses new data or his progress on experiments, he always asks his PI, “Is there anything else you wanted to talk about?”

Setting an agenda helps us to introduce uncomfortable topics. For example, including ‘summer course funding’ in her agenda helped Tess to request funding for a course on computational neuroscience — something she had been avoiding doing for weeks. It turned out that Tess’s PI was happy to provide support.

We and our PIs see our projects from different perspectives. Whereas they focus on the big picture, we wrestle with implementation. Because of this disconnect, we can discount their advice as being out of touch. Conversely, if we shoot down all their suggestions for ambitious experiments, our PIs grow frustrated.

When we realize we’re saying ‘no’, we try to engage with our PI’s idea by asking specific questions. These moments of potential conflict can turn into opportunities to hash out experimental strategies. We might say, “I think that would be an exciting direction, and it would be helpful for me if we could discuss specific metrics for measuring that result.” Instead of searching for flaws, we try to discuss a realistic road map for an optimistic outcome.

We are never going to be perfect mentees. We remind each other to take an active role in our mentoring relationships and to seek mentorship from multiple sources. Tess has great conversations with her physician–scientist PI about her clinical interests as an MD–PhD student. But she also has female mentors for advice about working within a male-dominated field. Samuel routinely discusses personal career goals with his PI, but relies on collaborators for advice on experimental techniques outside his PI’s expertise.

Discussions on mentorship often place the onus solely on the mentor. But, as mentees, we also need to ask ourselves, “What’s working and not working in this interaction? Where can I try something new? What would be ideal?” No template can solve all PI–student concerns. But simple steps can go a long way in helping these relationships to thrive.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

12 June 2019

Date