B. BACKGROUND, SIGNIFICANCE, and INNOVATION

Introduction to causal inference. Perception is best viewed as a problem of probabilistic inference, in which the brain infers the state of the environment from stochastic and incomplete patterns of sensory input (e.g., von Helmholtz, 1867; Yuille and Bulthoff, 1996; Lee and Mumford, 2003; Kersten et al., 2004; Fiser et al., 2010; Pouget et al., 2013; Haefner et al., 2016). Using this input, the brain builds and constantly updates an internal model of the world that guides action planning (Kawato and Wolpert, 1998). In the context of perception, "causal inference" refers to the process by which the brain adjudicates between different causes in the world that could produce similar patterns of sensory input (Kording et al., 2007; Shams and Beierholm, 2010).

As a model of causal inference, we propose to study perception of object motion and depth during self-motion. If an object moves across the retina's photoreceptor array over time ("image motion"), the brain must decide whether this pattern is produced by an object moving in the world, by the animal's eyes, head, or body moving, or by some combination of object and self-motion (e.g., Royden et al., 1994; Royden and Hildreth, 1996; Rushton and Warren, 2005; Warren and Rushton, 2009; Dokka et al., 2015; Dokka et al., 2019). In this context, the brain's goal is to infer the object's velocity and location in the world and the observer's self-motion, as these are the variables likely to be needed to guide action. The image motion of an object that is stationary in the world carries information about self-motion, while the image motion of a moving object carries information about object motion and self-motion. Thus the same pattern of image motion may be caused by two generative models for sensory inputs (Dokka et al., 2019); causal inference seeks to identify the most likely model. In experimental psychology and computational neuroscience, causal inference has been formulated in a normative Bayesian inference framework (Knill, 2007; Kording et al., 2007), which predicts human (e.g., Roach et al., 2006; Kording et al., 2007; Hospedales and Vijayakumar, 2009; de Winkel et al., 2017; Magnotti and Beauchamp, 2017; Perdreau et al., 2019) and animal (Dokka et al., 2019; Fang et al., 2019) behavior.

Our computational and conceptual approach. In neural population recordings and manipulations, we will test model-derived hypotheses for how and where critical variables of the causal-inference computations are represented in the brain, and how they evolve with time. This involves two closely intertwined stages of theoretical work. First, we will develop normative models that use Bayesian causal inference theory and/or real-time rational control theory, and can be fit to behavior to estimate latent variables involved in causal inference (**Project A**). Second, we will use a variety of cutting-edge data analyses to identify how and where these latent variables, as well as observable variables, are represented in the brain, and how they predict behavior over time (**Overall Strategy** & **Data Science Core**). This approach necessitates tight integration of experimental design and model development, which requires a coordinated multi-investigator effort.

A critical (and often overlooked) conceptual theme involves understanding how high-level signals related to categorical decisions about causes are used to update sensory representations. To maintain internal consistency with the animal's current beliefs about the structure of the environment, it has been proposed that sensory representations of task-relevant variables should be updated based on the organism's beliefs (Stocker and Simoncelli, 2008; Luu and Stocker, 2018). Based on this work and our preliminary data, our ***overall hypothesis*** *is that feedback connections from parietal and/or prefrontal regions update sensory representations in a coordinated fashion across brain areas to maintain consistency with the animal's current beliefs about the world*. This idea implies a **novel and critical role for feedback projections** in propagating beliefs about the state of the environment from decision-making regions back to sensory areas. To test this hypothesis, we will selectively manipulate feedback pathways during performance of perceptual decision-making tasks.

Theoretical terminology

Bayes rule: Equation describing how to update one's belief about an unobserved variable x after making observation o: $p(x|o) = p(o|x)\ p(x) / p(o)$.

Posterior, $p(x|o)$: Probability distribution representing knowledge about the unobserved variable x after having observed o.

Likelihood, $p(o|x)$: The probability of observing a particular value of o as a function of x (probability density if o is continuous).

Prior, $p(x)$: Probability distribution representing knowledge about the unobserved variable x before making an observation.

Causal inference: Special case of Bayesian inference in which inferences are made not just about the value of x but also the structure of the likelihood.

Latent variable: A variable that is not directly observed, but has to be inferred from observations using a model consisting of likelihood and prior.

Advances over prior research. Little is known about the neural circuits that mediate causal inference. A few human neuroimaging (fMRI, MEG, or EEG) studies identified neural correlates of causal inference in parietal and prefrontal cortex (Rohe and Noppeney, 2015, 2016; Cao et al., 2019; Rohe et al., 2019), without sufficient

spatial or temporal resolution to study the neural circuitry and temporal dynamics of the process. In a visually guided reaching task in non-human primates, neural activity in premotor cortex represents target location, consistent with a Bayesian causal inference model (Fang et al., 2019). However, this study did not require animals to report their causal inference, nor did it explore how other representations of task-relevant variables are updated according to the animal's beliefs about the states of the world.

Visual motion perception has provided a powerful model for understanding the neural basis of sensory discrimination and perceptual decision-making (Parker and Newsome, 1998; Gold and Shadlen, 2007). Yet the vast majority of previous studies are unnatural because the eyes, head, and body were constrained to remain stationary. Under such simplistic conditions, there is a unique mapping between object motion in the world and image motion on the retina. Our proposal addresses the more general situation in which the brain must use causal inference to determine whether the image motion results from self-motion, object motion, or both.

Because the perception of object motion and depth during self-motion involves judgments of multiple variables that are naturally related through the physical structure of the environment, rather than arbitrary trained-in associations among cues, our Bayesian observer models allow us to make specific quantitative predictions for how estimates of one variable should depend on current beliefs about the others (**Project A**). For example, if an observer incorrectly infers that a moving object is stationary in the world, then the component of image motion due to object motion must be "explained away" by the observer's estimate of their self-motion or object depth (**Fig. 1**).

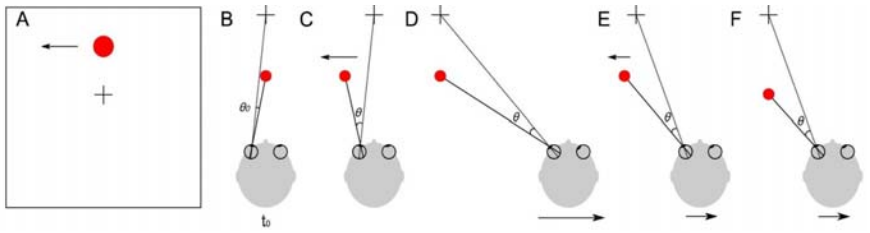

Figure 1. Tradeoffs between object motion, depth, and self-motion. (**A**) Image motion of object (red circle). (**B**) Starting position of an observer viewing the scene with one eye. (**C–E**) The image in panel A could be produced by leftward movement of the object in the world (C), rightward movement of the observer (D), or some combination of object motion and self-motion (E). If the observer fails to detect that the object is moving in the world in panel E, then they could "explain away" the extra image motion by perceiving the object to be at a different depth (as in F) or by perceiving their self-motion to be faster (as in D).

In a trial-based experimental design, we will test these model-driven hypotheses of how and where these variables are represented in the brain, and how they update based on changes in belief about whether an object moves in the world (**Project B**). This model system also supports more naturalistic behaviors. By studying a dynamic task that unfolds continuously in time, we will evaluate the behavioral and neural dynamics of how beliefs are represented and sensory representations are updated to maintain consistency with beliefs about object motion (**Project C**). The use of a common perceptual model system across all of our experimental studies and computational models ties together all of the projects, as well as allowing our theoretical frameworks to have a common structure of underlying variables, which facilitates extracting the critical computational motifs across species and tasks.

Finally, tackling the mechanisms of causal inference in the context of object motion, depth, and self-motion allows us to build on our extensive previous research on the neural computations underlying perception of depth, self-motion, and object motion separately. Using both correlative and causal manipulations, we discovered neural representations of depth from disparity and motion parallax cues (DeAngelis et al., 1998; DeAngelis and Newsome, 1999; Uka and DeAngelis, 2006; Chowdhury and DeAngelis, 2008; Nadler et al., 2008; Nadler et al., 2009; Nadler et al., 2013), representations of self-motion based on visual and vestibular signals (Gu et al., 2006; Gu et al., 2008; Chen et al., 2011; Fetsch et al., 2012; Chen et al., 2013), and neural correlates of computations that jointly represent object motion and self-motion (Sasaki et al., 2017, 2019). We also established, with recent behavioral studies in both monkeys and humans, that causal inference accounts nicely for perception of self-motion in the presence of object motion (Dokka et al., 2019). Moreover, our preliminary behavioral data (below) establish that this model system shows the diagnostic features of a causal inference process, including effects of optic flow on perceived object motion and interactions between perceived depth and object motion. Thus, we are now ideally positioned to tackle the problem of how the joint neural representations of object motion, depth, and self-motion are modulated by causal inference, and to identify specific neural pathways involved in these computations.

Broader scientific impacts. The proposed work will make four major contributions to neuroscience. (1) Results of these projects will establish the first direct neural correlates of categorical causal inference at the level of single neurons and neural populations. (2) Our studies will provide the first systematic investigation of

how beliefs about states of the world are propagated from decision-making regions back to sensory regions of the brain, which has important implications for understanding the functions of cortical feedback projections. Feedback projections are ubiquitous in the brain, and have been suggested to play a number of different roles in modulating sensory processing, including mediating effects of selective attention (reviewed by Paneri and Gregoriou, 2017; Knudsen, 2018), working memory (reviewed by Gazzaley and Nobre, 2012; Xu, 2017), prediction (Rao and Ballard, 1999), and prior beliefs (Fiser et al., 2010; Haefner et al., 2016). Our proposal suggests an important additional role of top-down feedback in updating sensory representations to be consistent with causal inferences, and our models will make predictions for the effects of causal manipulations of activity in specific feedback projections, which we will test. (3) Updating of sensory representations by top-down belief propagation also has important consequences for understanding the neural code in sensory areas of the brain. In particular, our framework predicts that some neurons in sensory areas represent posterior distributions over task variables rather than just the likelihood functions of sensory inputs. (4) Our experiments recognize that real-world causal inferences evolve dynamically over time, so that there is an important loop to close between perception and action. Actions influence sensory inputs, which then update beliefs about the causal structure of the world. We propose that these beliefs about causal structure in turn modulate representations of sensory variables via feedback connections to maintain internal consistency. (5) Health-related impact: Recent studies (Jardri and Deneve, 2013; Jardri et al., 2017; Noel et al., 2018) suggest that behavioral deficits in schizophrenia and autism spectrum disorders are linked to dysfunction of casual inference; thus, understanding the neural mechanisms of causal inference is critical for developing therapies.

Alignment with BRAIN Initiative goals. The proposed project addresses three of the seven major topic areas of the BRAIN 2025 strategic plan. The first area is “identifying fundamental principles.” Causal inference is a foundational principle of perception, and we propose the first systematic investigation of its neural substrates at cellular and circuit resolution. The principles that we uncover will be captured in computational models that will make predictions for new behavioral and neural manipulations. The second area is “demonstrating causality.” We will use multiple approaches to causally test the hypothesis that feedback projections from parietal and prefrontal cortex update sensory representations to align with inferences about causes of sensory signals. To rigorously establish causality, we will take advantage of optogenetic approaches in monkeys to selectively manipulate feedback pathways to sensory cortex while animals perform causal inference tasks. The third area is “from BRAIN initiative to the brain.” We will apply state-of-the-art approaches for large-scale population recordings and causal circuit manipulation to address a fundamental problem that has broad applicability across all sensory-motor systems. In addition, we will apply the latest theoretical and analytical approaches, some of which were developed with previous BRAIN initiative funding (to Drs. Angelaki and Pitkow), to relate neural activity to causal inference across time and in multiple brain areas.

Why now and why us? The time is right for a major effort to explore the neural basis of causal inference. Behavioral studies in humans show that performance in a variety of perceptual tasks is well explained by Bayesian causal inference models (e.g., Roach et al., 2006; Kording et al., 2007; Hospedales and Vijayakumar, 2009; de Winkel et al., 2017; Magnotti and Beauchamp, 2017; Perdreau et al., 2019), and we and others found behavioral signatures of causal inference in macaque monkeys (Dokka et al., 2019; Fang et al., 2019). Prefrontal and parietal areas in humans are suggested to represent causal inference in perceptual tasks (Rohe and Noppeney, 2015, 2016; Cao et al., 2019; Rohe et al., 2019). Finally, theoretical frameworks for causal inference are well developed (Kording et al., 2007; Shams and Beierholm, 2010), allowing us to extend them to dynamic navigation tasks in this project. Thus, despite considerable uncertainty about the neural circuits that mediate causal inference, sufficient pieces are in place to motivate a well-focused effort that combines theory, electrophysiology, and causal manipulations to link circuit elements to variables in the computational models.

Our team has expertise in a variety of areas that are necessary to accomplish our goals, which are too ambitious for a single laboratory to achieve. Based on our past accomplishments and preliminary data, we are well-positioned to develop the behavioral tasks, including dual-report tasks. This approach will allow us to reap the benefits of both traditional trial-based tasks and more naturalistic dynamic tasks, and thus to identify common motifs across tasks and species. Our theoretical expertise will enable us to develop computational approaches to fit various types of behavioral data with causal inference models, allowing us to estimate latent variables and identify their neural representations through population recordings. In addition, we have the experience in chemical and optogenetic manipulation of neural activity needed to successfully execute causal

manipulations that can identify key roles for feedback pathways in propagating beliefs about states of the world from decision-making areas back to sensory representations. Finally, all core members of the team have collaborated closely with other members on previous projects, providing evidence that we work well together.

C. APPROACH

C.1. Motivations for our approach to studying the neural basis of causal inference

C.1.1. Choice of animal models. Achieving our goals requires animals that can perform sophisticated behavioral tasks and animals that can support large-scale neuronal population recordings and causal manipulations of neural circuits. We propose to study macaques and mice, using multiple behavioral task designs, which would not be feasible in any one laboratory alone. A major strength of macaques is that we can train them to perform difficult behavioral tasks, such as the dual-report tasks proposed here. A major strength of mice is that we can obtain dense neural recordings across much of their brain to identify neural correlates of causal inference in a largely unbiased fashion and later (beyond the scope of this project) take advantage of cell-type-specific genetic manipulations for circuit analysis. Thus, our approach will make the best use of both model systems. Comparing results from continuous navigation tasks between macaques and mice, using our dynamic models, will reveal which mechanisms of causal inference are common across species.

C.1.2. Motivation for multiple task designs. Sensory discrimination and perceptual decision making were traditionally studied using trial-based tasks involving forced-choice decisions (Parker and Newsome, 1998; Gold and Shadlen, 2007). However, such tasks do not accurately represent real-world behavior, leading some neuroscientists to move toward more natural, continuous, and dynamic tasks (Pitkow and Angelaki, 2017; Gottlieb and Oudeyer, 2018; Juavinett et al., 2018; Najafi and Churchland, 2018). We propose to use both types of task designs, taking advantage of the strengths of each approach to address different questions.

Forced-choice psychophysical tasks have advantages for studying neural mechanisms of causal inference. First, they will allow us to compare our results with established behavioral tasks and phenomena, such as optic flow parsing (**Project B**, Aim 1), in which the perceived direction of object motion is influenced by background optic flow. This experiment will test whether causal inference can modulate low-level, seemingly automatic perceptual interactions. Second, our dual-report task, in which subjects report whether an object is moving or stationary in the world and separately report whether its depth is near or far, will let us condition our analysis on the causal belief and examine distributions of behavioral or neural responses to other variables (**Project B**, Aims 2 and 3). This experiment provides additional power that may be important for identifying subtle effects (see section C.2.7), which is not possible in continuous tasks for which the stimuli are different on every trial. By studying how causal inference modulates neural processing as task complexity varies, we will provide a general assessment of the ubiquity of the neural circuit mechanisms that mediate causal inference.

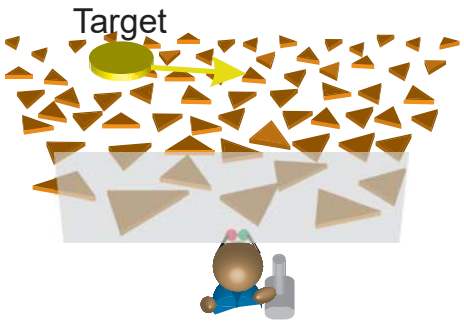

Figure 2. Causal inference 'firefly' task used in Project C. The animal sees a briefly presented target (yellow) that may or may not move in a virtual environment, and then navigates to the predicted location of the target in order to intercept it.

In natural behaviors, beliefs about the states of the world that produce sensory input evolve dynamically over time, so neural representations should be updated continuously while animals make causal inferences. To understand these dynamics, we will train both macaques and mice in a continuous causal-inference task (**Project C**). We will extend a task we recently developed (Lakshminarasimhan et al., 2018), in which animals navigate to the location of an object that was viewed only briefly, akin to the problem of catching a firefly. In our new version of the task (**Fig. 2**), this firefly stimulus will be either stationary or moving. If the animal infers that the firefly was stationary in the world, then the task is to navigate to its remembered location. On the other hand, if the firefly is believed to be moving, then the animal must extrapolate its motion to intercept the target at a future location. Thus, causal inference predicts how an observer's belief about the object's motion should govern the ability to accurately steer to the target.

By combining these task designs, this project will provide one of the first systematic comparisons of neural mechanisms across trial-based and dynamic tasks built around the same variables, which may help to resolve the longstanding tension between keeping experiments well-controlled and making them naturalistic. We will also take advantage of advances in normative modeling (Drugowitsch et al., 2014; Daptardar et al., 2019) to analyze more naturalistic tasks than previously possible.

C.1.3. Motivation for brain areas to be targeted in macaques. To test our central hypothesis that causal-inference signals from decision-making regions are fed back to update representations in sensory areas, we must record and manipulate neural activity in areas involved in causal inference, while simultaneously recording from relevant sensory representations. Human neuroimaging data suggest that parietal (Rohe and Noppeney, 2015, 2016) and prefrontal (Cao et al., 2019) cortex represent causal inference in perceptual tasks, so we selected parietal area 7a and prefrontal area 8av as candidates for causal inference about object motion in macaques, for the following reasons. Both 7a (Rozzi et al., 2006) and 8av (Petrides and Pandya, 2006) have direct feedback projections to areas MSTd and MT, which is necessary for the optogenetic approach proposed in Projects B and C. In addition, 7a and 8av have neural response properties suitable for representing causal inferences about object motion. Area 7a is involved in processing visual motion (Sakata et al., 1985; Read and Siegel, 1997; Siegel and Read, 1997; Phinney and Siegel, 2000; Merchant et al., 2004; Avila et al., 2018) and probably also in reference frame transformations (Andersen et al., 1990; Bremmer et al., 1997; Read and Siegel, 1997). Area 8av represents the direction and speed of moving dots (Hussar and Pasternak, 2009, 2013) and is involved in memory-guided perceptual decisions about motion (Hussar and Pasternak, 2012; Pasternak et al., 2015; Wimmer et al., 2016). Thus, the anatomical and physiological properties of these areas make them good candidates to propagate beliefs about object motion back to visual motion-processing areas.

We will record in area MT to examine the neural representation of object motion and depth during causal inference. MT represents motion (Born and Bradley, 2005), and our preliminary data (below) suggest that is involved in discounting self-motion to represent object motion. MT also represents depth from binocular disparity and motion parallax cues (DeAngelis and Uka, 2003; Uka and DeAngelis, 2003; Nadler et al., 2008; Nadler et al., 2013). Recordings will be made in area MSTd to examine how the representation of self-motion velocity is updated by causal inference. MSTd is involved in processing visual and vestibular signals related to self-motion (Britten, 2008; Angelaki et al., 2011); in addition, we and others have established causal links between MSTd activity and self-motion perception (Britten and van Wezel, 1998; Gu et al., 2012). Thus MT and MSTd are natural targets for studying how sensory representations are influenced by causal inference.

C.1.4. Motivation for theoretical frameworks. Studying the neural basis of computations in the brain requires assumptions about how these computations relate to its sensory inputs and behavioral outputs. One such assumption is (approximate) optimality—the idea that the brain's computations will be close to optimal in evolutionarily relevant contexts. Normative models yield predictions regarding the brain's beliefs about the outside world, and what its behavior *should* be. After such models are fitted to measured behavior, they can be used to look for neural correlates of unobservable (latent) variables in experimental data. Normative models are typically "algorithmic" (Marr, 1982); that is, they are constructed using invented variables like velocity or depth. In contrast, neural data such as spike times and counts describe a biological system ("implementation level" in Marr 1982). Whether, and if so how, the algorithmic and implementation levels are related is an empirical question that is answerable with statistical encoding and decoding methods (**Fig. 3**).

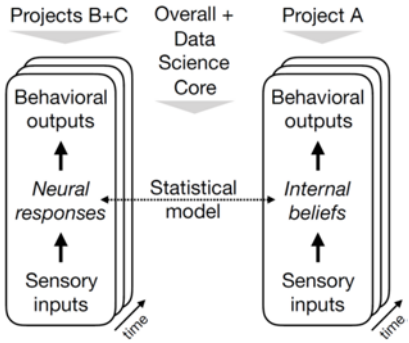

Figure 3. Core aspects of proposed theoretical approach to linking neural activity with internal beliefs.

Our theoretical framework will be constructed in three interacting steps: (1) build normative models of the brain's internal beliefs related to causal inference (**Project A**); (2) apply statistical encoding and decoding methods to neural activity to specify model parameters and manipulate neural activity to test these models (**Projects B, C**); and (3) develop an explicit encoding framework for linking neural responses to behavior (**Data Science Core, Aim 3**), and revise normative models iteratively as needed.

1) To build models that infer subjects' internal beliefs on a trial-by-trial basis, we will integrate two complementary approaches: an ideal observer model of the task (Brunton et al., 2013; Haefner et al., 2016; Tajima et al., 2016) and inverse rational control (Daptardar et al., 2019). Both approaches infer the task-relevant internal beliefs of subjects from their behavior, while assuming the brain to be rational with respect to the elements of the computations that are not constrained by the experiment.

2) To determine which neurons and neuronal populations represent internal beliefs, we will follow the prescriptions of Parker and Newsome (1998). In this step, we will determine how neurons change their responses as beliefs change (encoding), then we will verify that these changes contain the information

necessary to support behavioral performance (decoding), that response variability is correlated with behavior when external inputs are constant, and that causal manipulations of neural activity induce predicted changes in behavior. While Parker and Newsome (1998) formulated these criteria with respect to single neuron studies, we will apply them to populations concurrently recorded across multiple brain areas, as well as to circuit manipulations. We will also study within- and across-area communications between different neural populations representing these beliefs.

3) While decoding approaches to neural population analysis (e.g., linear decoders or naive Bayesian observers) have become standard, encoding approaches are still developing rapidly. We will combine three elements into a unified statistical encoding model of neural responses: first, hypothesis and observation-driven predictors (variables in our models, behavioral choices, stimuli); second, latent variables inferred in an unsupervised way (as in Gaussian process factor analysis); and third, variability in the timing of the internal computations across trials or time points (Duncker and Sahani, 2018). Although each element has been developed, combining them in a single method tailored to our tasks and data is non-trivial and a key goal (Aim 3) of our Data Science Core.

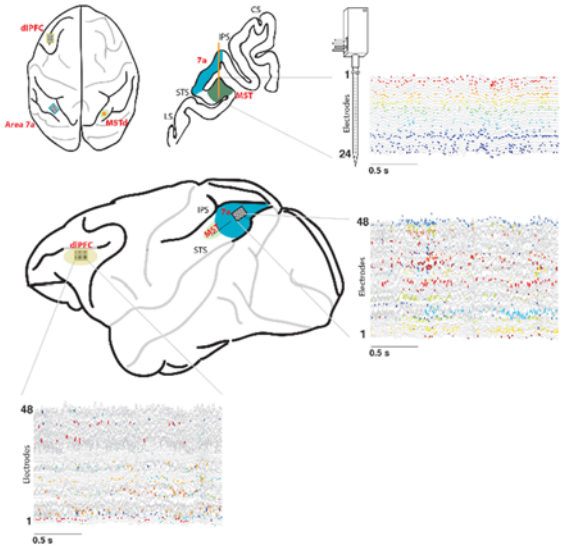

Figure 4. Large-scale parallel recordings in macaque monkeys. Top left: views of the macaque 7a (blue) and dorsolateral PFC (gold), as well as a linear electrode array inserted into area MSTd of the opposite hemisphere (green). Gray traces show analog signals, and colored dots indicate action potentials of single neurons.

C.2. Preliminary results

C.2.1. Monkey parallel recordings from multiple brain areas. A critical element of our experimental approach in macaques involves simultaneous recordings from parietal area 7a and prefrontal area 8av, which we hypothesize to be involved in causal inference based on human work (Rohe and Noppeney, 2015, 2016; Cao et al., 2019), and from visual motion processing areas MT and MSTd. In ongoing studies, we have made simultaneous recordings from Utah arrays in 7a and dorsolateral prefrontal cortex and from a 24-channel U-probe in area MSTd (**Fig. 4**). Our approach in Projects B and C will be similar, with two key modifications. First, the Utah array in prefrontal cortex will be shifted more posterior into area 8av, which has direct feedback projections to MSTd and MT (Petrides and Pandya, 2006). Second, we will use 64-channel U-Probes (and later NeuroPixels probes, once available and incorporated into our systems) to record from both areas MSTd and MT within a single vertical penetration. While moving to NeuroPixels probes will greatly increase our channel count and the richness of the data to be collected, our goals would be achievable using 64-channel U-probes.

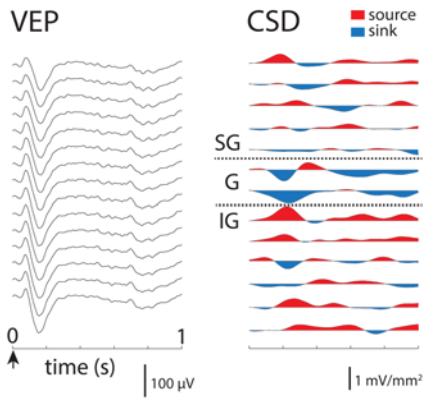

Figure 5. Laminar profile of field potentials (VEP, left) and current source density (CSD, right) recorded with a U-probe from area 7a. Arrow: stimulus onset. CSD reveals current sources (red) and sinks (blue), which are used to associate probe channels with granular (G), supragranular (SG), and infragranular (IG) layers (Schroeder et al. 1998).

C.2.2. Laminar recordings. The Angelaki and DeAngelis laboratories routinely record from neural populations using linear electrode arrays. By recording field potentials, we use current source density analysis (Schroeder et al., 1998) to identify patterns of current sources and sinks that correspond roughly to boundaries between supragranular, granular, and infragranular layers of cortex (**Fig. 5**). We will use this technique in Projects B and C to examine how neural correlates of variables in our models are represented across layers of cortex.

C.2.3. Monkey behavioral and neural correlates of flow parsing. We have collected preliminary data that demonstrate behavioral and neural correlates of flow parsing in macaque monkeys. Flow parsing is a visual mechanism by which background motion associated with self-motion is subtracted from retinal image motion to compute object motion in the world (Rushton and Warren, 2005; Warren and Rushton, 2007, 2008, 2009; Rushton et al., 2018). Monkeys discriminated the direction of motion (leftward vs. rightward of vertical) of a small patch of dots that was

surrounded by a background of optic flow to simulate forward, backward, or no self-motion (**Fig. 6A**, inset). Their behavior was biased by surrounding optic flow (**Fig. 6A**) as expected from human studies (e.g., Warren and Rushton, 2009), and this bias depended primarily on horizontal (but not vertical) location in the visual field (**Fig. 6B**). This location dependence is expected, as only the subtraction of horizontal components of optic flow should bias perceived direction in this task. To account for this behavior, we would predict that forward and backward optic flow have opposite effects on responses of MT neurons that prefer leftward and rightward motion (**Fig. 6C**), as we indeed observed (**Fig. 6D**). Single-session decoding of MT activity (using 32-channel Plexon V-probes) predicted perceptual biases that match well with behavior (**Fig. 6E**). By modeling MT responses as a weighted combination of tuning in retinal and world coordinates, we find that the representation of

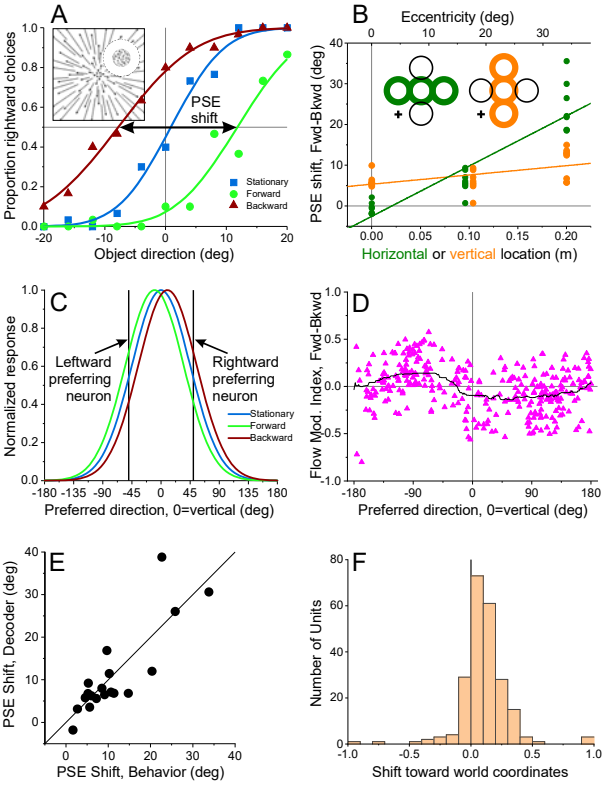

Figure 6. Preliminary data from flow-parsing task. (**A**) Background optic flow biases perceived direction of a patch of moving dots. (**B**) Biases depend on horizontal, but not vertical, patch location, as expected from flow-parsing mechanism. (**C**) Expected effect of optic flow on population activity in MT, to account for behavioral biases. The lateral shift predicts different effects on responses for neurons that prefer rightward vs. leftward motion (vertical lines). (**D**) Differential MT response between forward and backward optic flow depends on preferred direction, as expected from C. (**E**) Single-session decoding of MT activity predicts biases comparable to behavior. (**F**) Modeling MT responses reveals that tuning in MT is partially shifted toward world coordinates.

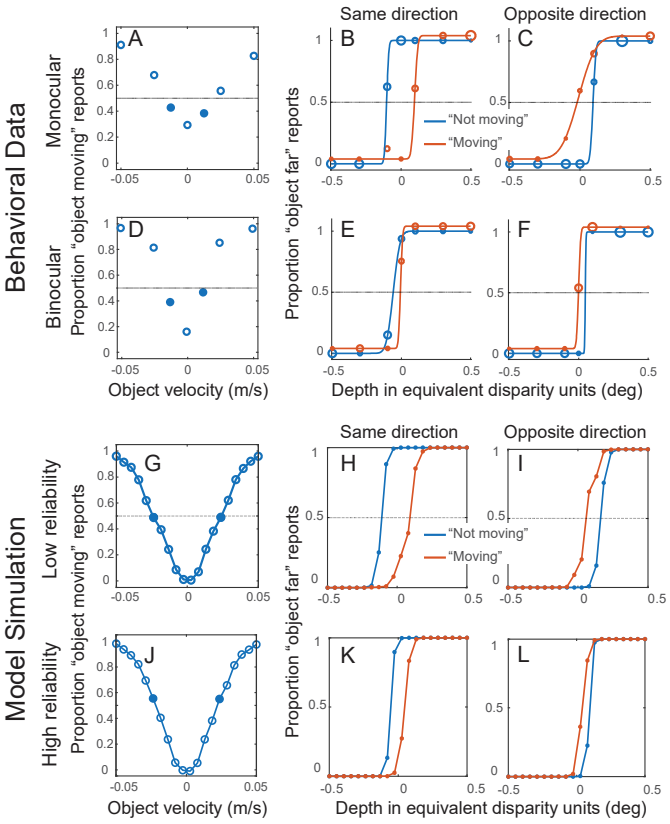

Figure 7. Data from a human (A-F) and model predictions (G-L) for the dual-report task to be used in Project B, Aims 2 and 3. (**A**) Proportion of "object moving" reports as a function of object velocity in the world. (**B,C**) Depth psychometric functions, sorted by report of object moving (red) or not moving (blue). Data are shown for object velocities indicated by filled symbols in A. In B, the object moves in the same direction as the observer; in C, they move in opposite directions. (**D-F**) Same as A-C, but for an object viewed binocularly, instead of monocularly. (**G-I**) Predictions of the Bayesian ideal observer model (see Fig. 1, Project A) for the case of low depth reliability, which is analogous to the monocular viewing situation of panels A-C. (**J-L**) Predictions of the model for the case of high depth reliability, which is analogous to the binocular viewing condition of panels D-F.

motion in MT is shifted slightly from retinal coordinates toward world coordinates (**Fig. 6F**).

Together these findings suggest that MT activity is modulated by signals related to optic flow to help represent object motion in world coordinates. However, in these experiments, it was clear that the object moved independently in the world; more generally, flow parsing should only occur when an object moves independently of the scene. Otherwise, image motion of the object should be integrated with that of other background elements to provide self-motion information. Aim 1 of Project B will examine whether these neural correlates of flow parsing are modulated by causal inference, thus testing whether it affects low-level sensory processes.

C.2.4. Tradeoffs of perceived depth and object motion in human behavior and a Bayesian observer model. We collected preliminary behavioral data from

humans (**Fig. 7**) in the dual-report task to be used in Project B (Aims 2 and 3). The observer experiences simulated lateral motion over a ground plane while viewing a scene consisting of several stationary objects, as well as one object that may or may not move independently (see Fig. 2 of Project B). Observers report whether this object was moving relative to the scene, and also whether it was near or far relative to the fixation target. As the speed of object motion increased, observers were more likely to report the object as moving (**Fig. 7A**). When we examined depth psychometric functions for object velocities at which subjects were about equally likely to report the object as moving or stationary, we found systematic biases in perceived depth contingent on the subject's belief about whether the object moved (**Fig. 7B,C**). When depth cues were made more reliable by adding stereo cues, these biases were substantially reduced in magnitude (**Fig. 7E,F**), consistent with the expectation that perceived self-motion velocity could explain away most of the retinal velocity associated with object motion in the world. Our Bayesian observer model (see Fig. 1 of Project A) accounts nicely for all of the main features of these preliminary behavioral data (**Fig. 7G-L**), *supporting our overall hypothesis that sensory representations of task-relevant variables are updated based on causal inference*. These findings motivate the neural population recordings and causal manipulations in Project B to test this hypothesis.

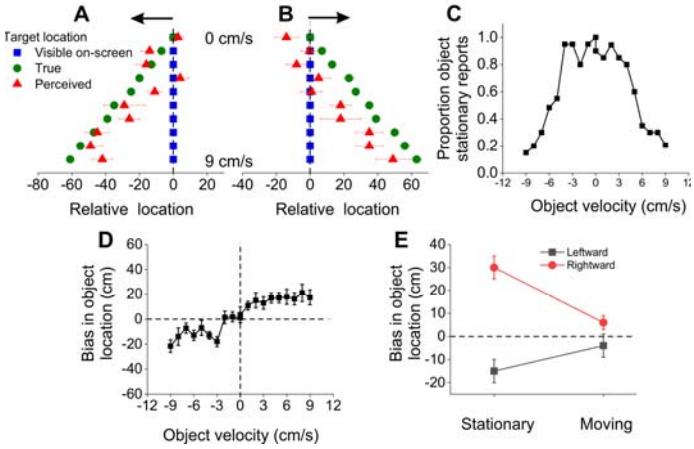

Figure 8. Preliminary data from a human subject for the dual-report firefly task. (A,B) Navigation behavior for different object speeds (collapsed along horizontal dimension only). Blue: object's location before it disappears; Green: true object location when subject stops; Red: subject's perceived object location. (**C**) Proportion of "object stationary" reports. **(D)** Bias in perceived object location as a function of object speed. (**E**) Bias conditioned upon belief about object motion for speeds that produced similar proportions of stationary vs. moving reports.

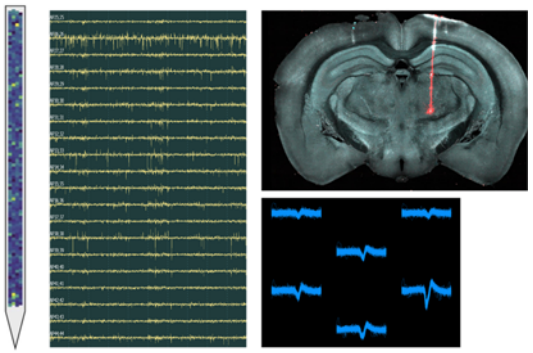

Figure 9. Mouse Neuropixels recordings. Left: Schematic of probe, with spike amplitudes color-coded (colder colors, smaller spikes, warmer colors, larger spikes). Middle: Traces from 20 electrodes, showing spiking on some channels. Top right: CM-DiI labeled fluorescence microscopy shows probe track. Channels span entire marked (red) area traversing entire cortex, hippocampus, and part of midbrain. Bottom right: waveforms of single units.

<u>C2.5. Navigation to a moving 'firefly' reflects causal inference regarding object motion</u>. Preliminary human data on the dual-report firefly task (**Fig. 8**) show that, for low object speeds, the subject reported the firefly to be stationary on almost every trial (**Fig. 8C**). In parallel, the stopping location was biased in a direction consistent with incomplete flow-parsing (**Fig. 8D**), in which the perceived velocity of object motion is influenced by background optic flow. When conditioned on the subject's report of object motion (considering only the two object speeds with roughly equal proportions of stationary vs moving reports), bias was consistently larger for stationary beliefs (**Fig. 8E**), as the subject failed to correctly incorporate object speed into their prediction of the future location of the target. We will exploit this belief-dependent steering bias to study the dynamic neural correlates of causal inference in macaques in Project C (Aim 2).

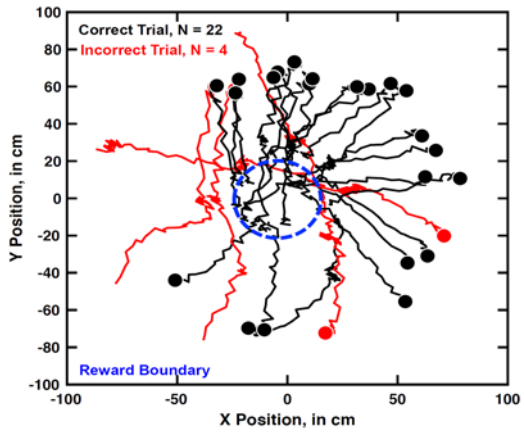

Figure 10. Data from a mouse performing the base version of the firefly task. In this diagram, the location of the remembered target is plotted at the origin, and circles denote different starting locations of the animal within the virtual environment. Black trajectories represented trials that ended within the reward zone (blue dashed circle); red trajectories indicate incorrect trials.

<u>C.2.6. Mouse NeuroPixels recordings.</u> The Angelaki laboratory is now using Neuropixels probes (Steinmetz et al., 2018; Juavinett et al., 2019) to record from large populations of neurons across much of the mouse brain (**Fig. 9**, see also Vertebrate Animals section for Project C). These 960-channel linear arrays (384 addressable

channels) have very finely spaced electrode contacts that allow for good unit isolation and densely sampled recordings. These probes will be used in Project C (Aim 3) to map out neural representations of observable and latent variables (from our models) across many regions of the mouse brain.

C.2.7. Mouse behavior in the continuous navigation task. We are training mice (**Fig. 10**) to perform the basic firefly task (without independent object motion) for Project C. In most trials, the animal correctly navigates to the location of the remembered firefly, ending within the reward zone. Thus, we do not anticipate difficulties with training mice to perform the causal-inference version, in which the firefly may move during each trial.

C.3. General Experimental Methods

C.3.1. Monkey preparation and behavioral training. Rhesus macaque (*macaca mulatta*) monkeys (age 3-10 years, weight 3-15 kg) will be used in Projects B and C. Animals will be surgically prepared by implantation of a head-restraint device and a recording grid and platform to allow multiple linear electrode arrays to be targeted to specific brain regions (see Vertebrate Animals for details). Some animals will be implanted with scleral coils for measuring eye movements, and others will undergo video-based eye tracking. Animals will then be trained to perform either trial-based psychophysical tasks (**Project B**) or continuous navigation tasks (**Project C**) using standard operant conditioning with positive liquid rewards. Animals will be allowed to work to satiety, while also satisfying institution-specific criteria for water scheduling (see Vertebrate Animals for details).

C.3.2. Mouse preparation and behavioral training. C57Bl/6J mice will sit comfortably, with head restrained, on a track ball that lets them walk or run to navigate through a virtual environment (see Vertebrate Animals for details). Each mouse will be surgically implanted with a head post for restraint in the virtual reality system. Animals will be trained to perform the firefly task (no dual report) using standard operant conditioning techniques, with positive fluid rewards. During training and data collection, water intake will be restricted, weight and excretions will be monitored daily, and supplemental fluid and/or fruit will be provided as necessary.

C.3.3. Common methods for neural population recordings in monkeys. Simultaneous chronic recordings will be obtained from areas 7a and 8av of one hemisphere using 128-channel Utah arrays (Mitz et al., 2017). Arrays will be implanted in a sterile surgery and inserted into cortex using a high-speed pneumatic inserter. Data from the two Utah arrays will be acquired with a 512-channel Ripple Grapevine system. We have already obtained simultaneous Utah array recordings from parietal and prefrontal cortex (**Fig. 4**) using these methods.

We will also record simultaneously from sensory representations using linear electrode arrays inserted into areas MT and MSTd in each session. We initially plan to record from MT and MSTd using a single 64-channel U-probe because we can target both areas in a single vertical penetration. If necessary, however, our systems will allow us to drive multiple independent U-probes. We have extensive experience recording from areas MSTd and MT with these types of arrays (**Figs. 4** and **6**, respectively). To increase the channel count of our recordings, we plan to transition to recording with NeuroPixels probes (Umec) in the second year of the project, when we anticipate versions of these probes suitable for monkey experiments will become available. Although these experiments do not allow precise cell-type identification, we should be able to infer general categories of putative excitatory and inhibitory neurons from spike waveforms.

C.3.4. Reversible inactivation of neural activity in monkeys. We will reversibly inactivate areas 7a and 8av on the side ipsilateral to our recordings in MT and MSTd by injecting the GABA agonist muscimol. Injections of muscimol (1-4 microliters each, 5-10 mg/ml) will be carried out using a “microinjectrode”, which combines a fine cannula and a microelectrode (see Vertebrate Animals for details). This approach allows neural activity to be monitored while drug is injected, to confirm that the drug suppresses activity. We have extensive experience with this technique in multiple brain areas (Chowdhury and DeAngelis, 2008; Gu et al., 2012; Chen et al., 2016). We have requested funds to purchase modified Plexon U-probes that contain ports for drug injection within the linear electrode array. This technology will improve our mapping of the inactivated regions.

C.3.5. Optogenetic manipulation of feedback pathways in monkeys. To test our overall hypothesis that feedback from 7a and/or 8av updates sensory representations in areas MT and MSTd, the ideal approach is to selectively manipulate activity in these feedback projections (El-Shamayleh et al., 2016). Non-selectively activating or inactivating neurons in 7a and 8av is less desirable, as it will affect multiple projections, not just the feedback projections to MT and MSTd. Thus, we propose to express opsins in 7a and 8av through injection of a viral vector, and to illuminate the axon terminals of 7a and 8av neurons that have direct feedback projections to MT and MSTd. This approach was recently used in monkeys to suppress feedback projections

from V2 to V1 (Nurminen et al., 2018), and one of our team members (Dr. Horwitz) used this approach in his laboratory to manipulate feedback projections from V1 to the superior colliculus (El-Shamayleh et al., 2016).

Our goal will be to suppress activity in feedback projections by injecting AAV9-CAG-ArchT-GFP into area 7a or 8av. ArchT is a proton pump, so activation of the axon terminals projecting from 7a/8av to MT or MSTd should suppress the feedback signals. With this optical approach, we will simultaneously record neural activity in areas MT and MSTd (see above and Project B) while feedback is inactivated. We expect that this manipulation will reduce or eliminate the updating of representations of depth (in MT) and self-motion velocity (in MSTd), as compared to interleaved trials without illumination of the axon terminals. This approach will allow us to identify specific neural pathways that are involved in causal inference.

If it proves difficult to get sufficient expression and efficacy of ArchT, we will inject AAV1-hSyn-ChR2(H134R)-mCherry into areas 7a and 8av to express channelrhodopsin (ChR2). This approach would allow us to activate the axon terminals of feedback projections from 7a/8av to areas MT and MSTd. Because the excitation produced would be non-specific, and certainly not matching the activity pattern that selectively updates sensory representations, we would also expect this manipulation to interfere with sensory updating during the causal inference tasks. We will also use injections of ChR2 into 7a/8av in preliminary experiments to confirm histologically that axon terminals in areas MT and MSTd are labeled.

C.4. General analytical approaches for relating neural activity to observable and latent variables

C.4.1. Rationale. Our neural analysis has multiple aims. First, we aim to identify the neural correlates of the various observed and latent variables that are relevant to the proposed tasks. Second, we aim to uncover inter-area communication that results from processing these variables. Third, we aim to determine how neural population dynamics deviate from those predicted by model-based latent variables, in order to refine our models and to close the theory-experiment loop. We will address these aims using multiple, complementary methods, including neural encoding, decoding, and state-space models.

C.4.2. Variables. Our normative models will predict, for each experimental trial, the values that task-relevant latent variables are expected to take. These estimates depend on model parameters that will be tuned to best match the observed behavior. The exact set of latent variables will depend on the task, and might include estimates of self-motion, distance to object, and whether an object is moving or stationary. As our normative models are probabilistic, they will also predict uncertainties associated with each of these estimates, which we will include as latent variables. We will additionally perform model-free neural analysis by replacing the model's estimates of some variables with their ground-truth values. For example, rather than using the model's estimated self-motion velocity, we would instead use the true self-motion velocity to guide our neural analysis. This ground truth is by design unknowable by the animals and so cannot be directly represented by their nervous systems. Nonetheless, it can act as a proxy for model-based estimates that, if close enough, can reveal the areas that dominantly encode these variables without the use of specific models. For model-based and model-free analyses, we will control for the neural representation of sensory and motor variables by including them in the analyses. Sensory variables include flow field velocity, retinal image velocity of an object, or a mouse's running speed. Motor variables concern all task-related behaviors, such as the monkey's decision reports, or the movement of the hand controlling the joystick. Including sensory and motor variables into the analysis will reveal their direct representation in the recorded neural populations.

C.4.3. Encoding and decoding approaches. We will use both encoding and decoding models to link model or task variables, $\boldsymbol{z}$, to neural population activity and dynamics, r. Encoding models predict neural responses given these variables, $p(r|z)$, and thus model how these variables are encoded. Decoding models predict the encoded variables from neural responses, $p(z|r)$, and thus aim to recover these variables from neural activity. Although an encoding model can be turned into a decoding model by Bayes' rule, $p(z|r) \propto p(r|z)p(z)$ (e.g., Park et al., 2014), each has its own benefits by encapsulating different assumptions about the neural code.

For individual neurons, we will use encoding models to characterize how the neurons are tuned to different variables. We will do so by averaging across trials in which these variables take similar values up to a certain precision (for Project B). For dynamic tasks, we will also study how tuning changes over time. We will use generalized linear models to characterize neural responses of individual neurons (Paninski, 2004; Truccolo et al., 2005) and populations (Pillow et al., 2008). For individual neurons, these models will include temporal filters for the different variables to describe the time course with which they influence each neuron's activity (Park et al., 2014), as well spike-history filters that account for effects like bursting and refractory periods.

Moving to neural populations, we will include neural coupling filters that allow us to determine how much variability can be explained by across-neuron coupling, and how much is driven by the variables of interest (Pillow et al., 2008). Overall, this approach will reveal which variables drive responses of the recorded neurons.

We will use multiple approaches to decoding variables from recorded populations. First, we will use linear decoders to simultaneously decode variables of interest. For dynamic tasks, we will ask if the decoder changes over time by training separate decoders at different points in time (e.g., Rigotti et al., 2013), and comparing their performance across time points. This analysis will identify linear subspaces of neural population activity that represent the different variables. Second, we will invert trained general linear encoding models and use them as decoders (Pillow et al., 2008; Park et al., 2014). Unlike linear decoders, these models rely on point processes to model individual spikes, but are also more difficult to train. Third, we will use canonical correlation analysis (CCA, Hardle and Simar, 2015) to simultaneously identify linear subspaces in neural population activity and task variables of interest that have maximum correlation with each other. In contrast to linear decoding, this analysis can identify combinations of variables that drive neural variability (in that sense, it lies between encoding and decoding models). Fourth, we will use a new artificial neural network-based method to identify probabilistic representations in neural populations (Walker et al., 2018). This method uses a non-linear decoder to decode full likelihood functions in individual trials, and has the potential to reveal details about uncertainty coding beyond the latent variables related to uncertainty in our normative models.

C.4.4. State-space models. These models are commonly used to identify the time-course of latent variables that best explain neural population activity (e.g., Smith and Brown, 2003; Yu et al., 2009). Their power lies in allowing us to specify only a model for how the latent variables evolve over time and how they map into higher-dimensional population activity, without having to specify the latent variables themselves. Thus, these models can unravel latent variables in individual trials without explicitly modeling them. We will extend these models to include the time course of predicted variables, effectively assuming that neural population activity is generated by a combination of modeled and non-modeled latent variables. Inverting this model, in turn, will allow us to identify the state of latent variables that we do not explicitly model. We will inspect these variables to inform us about residuals that our normative models do not capture, which may help to refine those models. Therefore, state-space models help close the loop between theory and experiment. See Data Science Core for details.

C.4.5. Across-area comparisons. We will use both encoding and decoding approaches to compare representations across brain areas and layers. In general, we will ask how well different variables are encoded or can be decoded to identify if individual brain areas preferentially code for specific variables. Encoding models tell us how well neural variability is driven by different variables, and thus inform us about preferred representations. Decoding models tell us how well we can recover different variables from neural activity, and thus also inform us about encoded information. A potential pitfall is that we might not record the same number of neurons from areas to be compared, or that information may be coded in a more distributed manner across areas. Thus, obtaining convergent evidence from multiple approaches is important to reach robust conclusions.

C.4.6. Across-area communication. We will use communication subspace analysis (Semedo et al., 2019) for model-free identification of linear co-variability of population activity across brain areas. We will then compare the identified subspaces with those identified by CCA in relation to encoded variables; overlaps of these subspaces can indicate the communication of specific task-related variables. This approach will support the identification of linear encoding and covariability, but does not consider that some task-related variables may interact non-linearly. To identify such non-linear communication, we will rely on our normative models, which can predict these interactions. Furthermore, our across-area comparison will identify area-specific representations of variables. Together, the two approaches can predict how different areas should interact non-linearly, which we will use to identify these non-linear interactions in neural population activity.

C.5. Research Team

Our team brings together a unique constellation of expertise that is ideally suited to tackling this major problem, including extensive experience with behavioral task design, large-scale electrophysiology, reversible inactivation, optogenetic manipulation of specific neural pathways, theory development, and state-of-the-art neural analyses. The team includes the following key personnel:

Gregory DeAngelis, University of Rochester (Lead PI; PI, Project B; PI, Administrative Core): Dr. DeAngelis is an expert on neural mechanisms of depth perception, motion perception, and multisensory

integration. He brings extensive experience with developing challenging behavioral tasks for monkeys, performing electrophysiological and reversible inactivation experiments, and relating neural activity to behavior.

Dora Angelaki, New York University (PI, Project C): Dr. Angelaki is an expert on multisensory coding and dynamic neural computations, applied to the vestibular and visual systems and to action-perception loops. She provides expertise on macaque neurophysiology across multiple brain areas, plus stimulation and inactivation causal manipulations. Her laboratory is one of the few comparing identical tasks in macaques and rodents.

Jan Drugowitsch, Harvard University (co-PI, Project A; PI, Data Science Core): Dr. Drugowitsch is an expert on normative models of decision-making, speed-accuracy trade-off, and decisions based on dynamically evolving evidence. He brings strong experience in design and fit of dynamic decision-making models to experimental data, and in Bayesian decision and control theory for neural and behavior modeling.

Ralf Haefner, University of Rochester (co-PI, Project A): Dr. Haefner is an expert on the neural basis of hierarchical probabilistic inference in the context of perceptual decision-making. He provides extensive experience in analyzing neural population data from awake behaving monkeys, and in using them to test the predictions of normative probabilistic models.

Xaq Pitkow, Rice University (collaborator on Projects A and C): Dr. Pitkow is an expert in theoretical and computational approaches to understanding and modeling neural representations, as well as applying cutting-edge analytical approaches to neural data. He brings specific expertise on using rational control theory to model continuous dynamic behaviors such as the firefly task of Project C.

Gregory Horwitz, University of Washington (collaborator on Projects B and C): Dr. Horwitz is an expert on visual processing, including color vision, eye movements, and perceptual decision-making. He is also a leader in developing optogenetic approaches for monkeys. He brings to this project experience with optogenetically manipulating specific neural projections, which will be a critical approach in Projects B and C.

C.6. Project Timetable

Most Aims of the Projects will begin immediately, whereas some Aims will commence later once other objectives have been met. For example, work on the optogenetic approaches will begin in Year 2 after animals are trained to perform the required tasks and neural recordings have commenced. The table below outlines the overall timeline for the research program.

		Year 1	Year 2	Year 3	Year 4	Year 5
Project A: Theoretical framework **Lead:** Haefner **Co-Lead**: Drugowitsch	Aim 1: Develop a causal inference model for perception of motion and depth during self-motion					
	Aim 2: Develop a model of dynamical causal inference for perception and control					
Project B: Causal inference in trial-based tasks **Lead:** DeAngelis	Aim 1: Determine how causal inference affects optic flow parsing					
	Aim 2: Elucidate neural correlates of causal inference and belief propagation					
	Aim 3. Use neural inactivation to identify circuits for causal inference and belief propagation					
Project C: Causal inference in continuous navigation tasks **Lead:** Angelaki	Aim 1: Develop an object/self-motion version of our dynamic navigation task in monkeys and mice					
	Aim 2: Use neurophysiology and perturbation to identify task variable representations in monkeys					
	Aim 3. Map neural activity throughout the mouse brain using parallel recordings with Neuropixels					
Administrative Core **Lead:** DeAngelis	Aim 1: Oversee and evaluate research progress					
	Aim 2: Facilitate team collaboration and communication					
	Aim 3. Coordinate information sharing...					
Data Science Core **Lead:** Drugowitsch **Co-Lead**: Haefner	Aim 1: Establish computational infrastructure for data…					
	Aim 2: Define, implement… data-preprocessing pipeline					
	Aim 3. …methods to track latent dynamic variables…					
	Aim 4. Data Sharing Plan: share data and code…					