The 11th CiNet International Conference Abstracts

*Subject to change.

Slot 1: Andrew Lampinen

Affiliation: Anthropic
Department / Section: Research
Job Title: Member of Technical Staff

Title: What can neuroscience and AI learn from each other? Case studies on generalization, and rational analysis

Abstract: How should neuroscience adapt to developments n AI, and what could AI still learn from neuroscience? In this talk, I'll first describe focus on what neuroscience can learn from AI, and advocate for the value of rich, naturalistic experimental paradigms and complex models. I'll describe how these can provide a route to more generalizable theories through methods like rational analysis. I'll then describe why I think AI has ***not*** been strongly influenced by neuroscience — despite the best efforts of many. I'll illustrate these ideas with two case studies from our research. The first focuses on content effects on logical reasoning in humans and language models, and how their similar patterns of behavior may originate in similar pressures. The second case study focuses on surprising findings about generalizations language models fail to make from their training data (such as reversals), and how these patterns of generalization are qualitatively different in context — and how that might connect to the role of hippocampal replay in the brain.

Slot 2: Wilma Bainbridge

Affiliation: University of Chicago
Department / Section: Department of Psychology
Job Title: Associate Professor

Title: Artificial intelligence can predict our memories better than we can

Abstract: Despite our unique individual differences, there's a surprising consistency across people in their memories, where we tend to remember and forget the same items. In other words, some images, sounds, words, and movements are more memorable than others, across observers. We have found that this effect is so pervasive that neural networks can predict people's memories based on images alone. In this talk, I will demonstrate how predictable people's memories are, even in real-world scenarios like remembering pieces from an art museum (Davis & Bainbridge, 2023), where AI outperforms humans in predicting the art we will remember (Davis et al., accepted). I will also show that AI can predict memory in children as young as 4 years of age (Guo & Bainbridge, 2024), and that its predictiveness can be used to measure the maturation of the developing brain. These results show support for our central hypothesis that the brain prioritizes easy to process items for memory (Bainbridge et al., 2025), and visual AI systems may pick up on this priority signal as well.

Slot 3: Kenji Doya

Affiliation: Okinawa Institute of Science and Technology
Department / Section: Neural Computation Unit
Job Title: Professor

Title: Toward more natural artificial intelligence

Abstract: The success of deep learning and foundation models has demonstrated that context-dependent, compositional representation learning is possible by error gradient descent, at the cost of huge training data and computation. While large language models achieved super-human knowledge and skills in words and codes, behaving in the physical world is still a challenge; child-level skills like folding a towel or opening a door are showcased as cutting-edge achievements. Under the IT industry's practice of first providing free or below-cost service to grab the market and then to exploit the dominance to raise money, heated competition for performance is posing serious problems in environmental and social sustainability. What do we need for building AIs that are more data and energy efficient learning and real-time adaptation in dynamic physical and social environments to better serve the humanity?

In this talk we seek insights from the brain and human societies for building more efficient, adaptive and safer AIs. Topics will include: 1) Parallel growing architectures with heterogenous learning algorithms and induction biases. 2) Active interactions in open environments with evolvable goals. 3) Social awareness and practices to avoid harms from monopoly.

Slot 4: John Duncan

Affiliation: University of Cambridge
Department / Section: MRC Cognition and Brain Sciences Unit
Job Title: Research Programme Leader

Title: The physiology of cognitive control

Abstract: In this talk I shall discuss the physiology of cognitive control. One classical approach seeks abstract control functions (e.g. inhibition, set switching), perhaps associated with specific coarse regions of the frontal lobe. I suggest that this way of posing the problem is mistaken. Behaviour is directed by a whole-brain model of the current situation, with many parts integrated into their correct roles and relationships. At the heart of this integration is a multiple-demand (MD) system well known from human brain imaging, with components closely connected, but widely distributed across the cortex. To address physiology, I shall discuss neural data from monkeys solving problems in an on-screen maze. With recordings from four putative homologues to human MD regions, in different regions of the frontal lobe, we examine key components of a mental model – current state, goal state, actions diminishing the difference between these two, and abstract problem structure. Across regions there was overlap but also wide quantitative variation in encoding fundamental task features. Sensory input and current state were strongly coded in ventrolateral frontal cortex, goal most stably in dorsomedial frontal cortex, and move most rapidly in both ventrolateral and dorsal premotor cortex.
Insula/orbitofrontal cortex responded during revision of a prepared route. Partial specialisations likely arise from distinct connectivity, while widespread copying of task representations reflects strong information exchange. Classical approaches to control, I suggest, should be discarded in favour of a whole-brain approach to construction and use of mental models.

Slot 5: Martin Schrimpf

Affiliation: Swiss Federal Institute of Technology in Lausanne
Department / Section: NeuroAI Lab
Job Title: Assistant Professor

Title: NeuroAI to Understand and Treat the Brain

Abstract: How can we understand the neural mechanisms underlying intelligence? I will argue that we are entering an era of digital brain models that are making great strides in forming an answer: these artificial neural networks not only reproduce human behaviors, but also do so via internally consistent representations. First, I will show that AI models optimized for real-world tasks, such as video understanding and language modeling, converge on strikingly brain-like solutions, which we quantify at scale through the Brain-Score community platform. But this convergence holds only up to a point: task optimization alone has begun to plateau, capturing roughly half of the explainable neural signal. I will make the case that the next leap will come from scaling neural data and optimizing models directly with large-scale brain recordings which will lead to next-generation multimodal predictive brain models. Finally, I will show how the best models already let us move from understanding to intervention: guiding stimulation to induce visual percepts via topography, and designing NeuroAI-based treatments such as fonts that improve reading in dyslexia. I will close on where NeuroAI might go next.

Slot 6: Takufumi Yanagisawa

Affiliation: The University of Osaka
Department / Section: Department of Neuroinformatics, Graduate School of Medicine
Job Title: Professor

Title: Exploring Representational Spaces for Visual-Imagery Brain–Machine Interfaces

Abstract: Recent advances in deep learning have enabled increasingly accurate decoding of perceptual, motor, and linguistic information from neural activity. In particular, embedding diverse external information—such as images and language—into a unified AI latent space has made it possible to decode neural activity even for previously unseen stimuli. Building on this, closed-loop brain–machine interfaces (BMIs) can now map a person's imagined visual content from neural activity onto a corresponding AI latent representation, and use this mapping to retrieve and present matching images from large-scale databases. Such systems exemplify Brain–AI integration: the joint use of neural and AI representations for information decoding and generation.

What AI representational space best maximizes the performance of this integration, and enables a practical visual-imagery BMI? We show that the structure of the AI representation shapes how well decoders trained on perceived images generalize to imagined visual content. In this talk, I will discuss how the choice of brain representational space affects Brain–AI integration performance, drawing on our recent findings.

I will also introduce our broader effort to translate these techniques into medical applications—communication for patients with severe motor impairment—by integrating implantable device development, neural decoding, output generation, and external device control into a unified BMI system.

Slot 7: Katharina Dobs

Affiliation: Justus Liebig University Giessen
Department / Section: Department of Computer Science
Job Title: Professor

Title: Functional Specialization in Biological and Artificial Visual Systems

Abstract: The human visual system contains striking examples of functional specialization, including category-selective regions that respond preferentially to faces, bodies, places, and objects. Does such specialization reflect biological constraints, or does it emerge as a more general principle of visual intelligence? In this talk, I will discuss how artificial neural networks can be used to investigate the origins and functional role of specialization.

I will present work showing that functional specialization can emerge spontaneously in deep neural networks trained for visual recognition, despite the absence of many biological constraints. I will further discuss evidence that both artificial and biological visual systems exhibit specialized as well as shared representations, highlighting the interplay between segregation and integration.

Finally, I will compare how category selectivity is organized across biological and artificial visual systems. While current deep neural networks capture important aspects of cortical organization, fundamental differences remain. Together, these findings illustrate how comparisons between brains and artificial neural networks can reveal both shared computational principles and important challenges for future models of visual intelligence.

Slot 8: Llion Jones & Kai Arulkumaran

Affiliation: Sakana AI
Job Title: CTO(Llion Jones), Research Scientist(Kai Arulkumaran)

Title: Bridging the gap between Neuro-science and Machine Learning

Abstract: Despite state of the art artificial intelligence supposedly being based on the brain with ‘artificial neural networks’ now driving the entire industry, there is surprisingly little overlap with modern neuro-science. ML is still using a neuron model from the 70s. Why is this, and what can we do about it?

Slot 9: Katherine Storrs

Affiliation: University of Auckland
Department / Section: School of Psychology
Job Title: Senior Lecturer

Title: Using the toolkit of psychology to evaluate visual understanding in artificial models

Abstract: Multimodal large language models (MLLMs) show remarkable performance in visual reasoning tasks, successfully tackling college-level challenges that seem to require a deep understanding of images. More recently, video-generating MLLMs have been claimed as "visual foundation models," implicitly learning the rules of causality, physics, and optical appearance. However, engineering benchmarks often evaluate performance on complex tasks rather than profiling a model's specific visual abilities, and few enable a direct comparison between human and model

performance. I will discuss three projects in which we use the toolkit of experimental and neuro-psychology to assess visual understanding in MLLMs. When testing MLLMs on a battery of 51 psychological assessments, we find they excel at high-level object recognition tasks while showing clinically-significant deficits on tests of low- and mid-level vision. Likewise, video-generating MLLMs perform well below humans on clinically-validated drawing tasks (e.g. copying a geometric figure), and at intuitive physics tasks (e.g. predict whether a tower of blocks will fall over). The visual deficits in current MLLMs demonstrate an example of "Moravec's Paradox" within perception: an artificial system can achieve complex object recognition without developing foundational visual concepts that in humans require no explicit training.

Slot 10: Robert Mok

Affiliation: Center for Information and Neural Networks (CiNet) / The University of Osaka / National Institute of Information and Communications Technology (NICT)

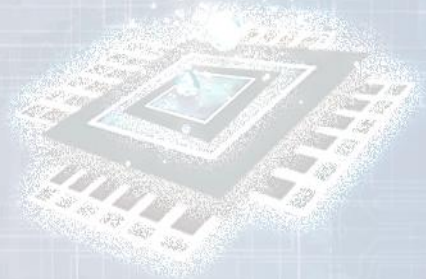
Job Title: Researcher (Tenure-Track) and Guest Associate Professor

Title: From neural codes to connectivity: how circuit organization shapes fast learning and generalization in the hippocampus

Abstract: The hippocampus is crucial for fast learning. Our ability to learn quickly and form one-shot memories is commonly attributed to the sparse code of the dentate gyrus (DG), which enables pattern separation and thereby facilitates fast learning. Yet this has largely been assumed rather than demonstrated, and we lack a principled account of why and when sparse coding helps or hurts learning, or what is traded off along the sparse-to-dense continuum. Does the degree of sparsity control the speed of learning, and must sparsity be implemented early in the processing stream, as in the DG, the input stage of the hippocampal trisynaptic circuit?

In this talk, I will present a systematic characterization of sparse coding in artificial neural networks trained on complex nonlinear tasks. Varying the level and position of sparsity, we find that both determine learning speed and the memorization–generalization trade-off. Learning speed follows an inverted-U function of sparsity, peaking at sparse codes that fall within the DG's estimated range. Critically, this benefit requires sparsity to be imposed early, mirroring the position of the DG in the trisynaptic circuit. Sparsity accelerates learning by producing high-dimensional, pattern-separated representations. But this comes at a cost: models learn fast by memorizing the inputs; they exhibit catastrophic failure on new examples, indicating that the DG's sparse code acts as an episodic snapshot mechanism rather than a system for extracting general structure.

However, the hippocampus is also implicated in learning abstract, generalizable knowledge. I will close with recent work showing that adding anatomically motivated connections – "skip connections" from entorhinal cortex to CA3 (skipping the DG) and recurrence – restores generalization without sacrificing learning speed.



Slot 11: Tatsuya Haga

Affiliation: Center for Information and Neural Networks (CiNet) /
National Institute of Information and Communications Technology (NICT)
Job Title: Researcher (Tenure-Track)

Title: Representations and dynamics in the hippocampus: Bridging space with concepts, and replay with sampling

Abstract: The hippocampal-entorhinal system has two major computational roles in the brain: representing cognitive maps and supporting memory. The hippocampus and entorhinal cortex exhibit spatial and conceptual representations, including place cells, grid cells, and concept cells, while the hippocampus replays past activity patterns to support efficient learning and decision-making. In this talk, I present two modeling studies related to these hippocampal functions. First, I present a unified neural representation model for space and concepts. The model produces biologically plausible hippocampal representations and enables analogical inference across both spatial structures and linguistic concepts within a single computational framework. Our results suggest that computational mechanisms suited for representing spatial information in the hippocampal-entorhinal system can be naturally extended to the learning and processing of semantic concepts. Second, I show that extending a Hopfield-type attractor neural network with momentum and kinetic energy provides a theoretical link between the dynamics and functions of hippocampal replay. This extension generates sequential memory recall resembling hippocampal replay and allows replay dynamics to be interpreted as Hamiltonian sampling in an augmented state space. Based on this interpretation, prioritized experience replay for reinforcement learning can be implemented through attractor dynamics. The model thus provides a computational account of how hippocampal dynamics can support efficient memory sampling.

Slot 12: Ida Momennejad

Affiliation: Microsoft Research NYC
Department / Section: Machine Learning/Reinforcement Learning
Job Title: Principal Researcher

Title: The Compositional Geometry of Flexible Cognition

Abstract: How are memories organized to enable both remembering the past and predicting the future? How does the latent geometry of brains and artificial neural networks enable compositional generalization and reasoning?

Flexible cognition moves continuously between cognitive modes, from memory to generalization, and from cached inference to multi-step reasoning and planning. I construe this flexibility as a walk over latent operations in the neural spaces of brains and neural networks. In human behavioral, neural, and modeling studies, we show that prefrontal cortex (PFC) and hippocampal hierarchies structure knowledge relationally, supporting remembering, prediction, and planning across scales. However, predictive cognitive maps do not address how operations emerge in the neural space and are composed into diverse cognitive functions. First, we built a recurrent agentic architecture for reasoning and planning, inspired by the PFC, with orchestration, action proposal, and evaluation as independent modules. Second, applying computational neuroscience methods to AI evaluation and interpretability, we find that the latent geometry of transformers supports some compositional generalization, and we trace the algorithmic geometry of primitives and their composition in reasoning models. This allows us to translate LLM reasoning traces into a graph that captures a walk over operations in the latent space, revealing inefficiencies and loops. This primitive transition graph offers insights and new objectives for continual learning, such as what we call Next Primitive Prediction (NPP), and paths toward training better reasoning models. I

conclude with a compositional framework for open-ended intelligence and how it informs PFC models.

Slot 13: Taro Toyoizumi

Affiliation: RIKEN
Department / Section: Center for Brain Science
Job Title: Team Director

Title: Modeling how the brain learns to represent ambiguous environments: Abstraction and Probability

Abstract: Adaptive behavior relies on activity-dependent synaptic plasticity to sculpt internal models of an uncertain world. I introduce two complementary frameworks for how the brain encodes abstraction and probability. Regarding abstract representations, we first propose a three-factor plasticity rule for nonlinear dimensionality reduction in a three-layer network inspired by the Drosophila olfactory circuit. This rule approximates the t-SNE algorithm and reproduces experimental findings from fly studies. Next, we describe a dual-pathway hippocampal model—featuring a dense, direct input path and a sparse, indirect input path—where modulation of inhibitory tone toggles recall between abstract categories and concrete exemplars. Regarding probabilistic representations, we exploit chaotic fluctuations in a recurrent network to perform Bayesian posterior sampling. Trained with biologically plausible learning on a cue-integration task, the network reliably approximates target distributions despite chaos-induced sensitivity to initial conditions. Together, these models illustrate how synaptic plasticity and neural dynamics could underlie abstract and probabilistic internal representations.

Slot 14
Speaker 1 : Lis Kanashiro Pereira

Affiliation: Center for Information and Neural Networks (CiNet),
National Institute of Information and Communications Technology (NICT)
Department / Section: Kitazawa Group, Neural Information Engineering Laboratory
Job Title: Senior Researcher

Title: Do LLMs and Brains Find the Same Things Funny (Omoroi)? Comparing Humor Representations in AI Models and the Human Brain

Abstract: While large language models (LLMs) excel across a wide range of complex tasks, their capacity for creative cognition, particularly humor processing, remains poorly understood. In this talk, we compare representations of humor in LLMs and the human brain using a Japanese Oogiri dataset with human funniness ratings. First, we compute a diverse set of computational features associated with humor, spanning basic linguistic properties, semantic and surprisal-based measures, and LLM-derived higher-order features such as ambiguity, incongruity resolution, and perspective shift. We then test whether these features predict human brain activity measured with fMRI during humor appreciation. Second, using mechanistic interpretability, we trace how humor-related information is represented and transformed across LLM layers. Together, our findings reveal where humor processing in LLMs converges with and diverges from the human brain at both the feature and unit levels.

Speaker 2: Takuto Kinoshita-Yamamoto

Affiliation: The University of Osaka; Osaka Neurological Institute
Department/Section: Department of Brain Physiology, Graduate School of Medicine; Department of Neurosurgery
Job Title: Neurosurgery Resident , PhD Student

Title: Human-Like Attention Emerges without Gaze Supervision: Three DINO-ViT Attention Systems and Their Neural Counterparts

Abstract: Humans learn where to look without being told. Strikingly, Vision Transformers trained by DINO—self-distillation without labels or gaze supervision, with an objective that effectively promotes output information maximization—develop attention patterns remarkably similar to human gaze. In our previous study, this human-like attention emerged specifically in DINO, not conventionally supervised ViTs. Even more unexpectedly, later-layer attention heads spontaneously separated into three functional groups: G1, selecting key locations within figures; G2, representing the whole figure; and G3, representing the background. Thus, a classical figure–ground dichotomy expanded into a three-component architecture, with G1 most closely matching human gaze.

We then asked whether these emergent artificial attention systems have counterparts in the human brain. Using 7T fMRI responses from eight participants in the Natural Scenes Dataset to 9,000–10,000 static natural images presented to each participant, we built stacked encoding models from G1, G2, and G3 features. The most surprising result was G1: despite the absence of motion, G1 strongly mapped onto motion-related cortex, including V4t, MT, MST, FST, and TPOJ, and extended into object/face cortex and FEF despite enforced fixation. G3 mapped preferentially onto medial scene/background cortex, while G2 occupied intervening territory. These results suggest that self-supervised information maximization may spontaneously discover organizing principles shared by artificial and biological vision.

Slot 15: Kentaro Abe

Affiliation: Tohoku University
Department / Section: Graduate School of Life Sciences
Job Title: Professor

Title: Language models for decoding non-human languages

Abstract: How language-related information is encoded and decoded in human brain remains to be revealed. Recent advances in large language models have demonstrated a remarkable ability to generate and reconstruct natural human language, a capacity once considered uniquely human. However, the mechanism by which biological brains and artificial language models arrive at similarly structured outputs remain unclear. To address this question, we use songbirds as a model system to investigate the neural mechanisms underlying language-like information processing. Birdsong consists of complex sequences of vocal elements governed by species- and individual-specific structural rules, making it a useful experimental system for studying how sequential information is generated, perceived, and represented in the brain. We have developed language models capable of learning the statistical and contextual structure of birdsong and generating naturalistic song sequences. In this lecture, I will introduce our approach to modeling and generating birdsongs using language models. I will also describe how the internal representations learned by these models can be compared with behavioural and neural data. By

combining computational modelling with neuronal recordings, we aim to examine how information about song sequences and communicative context is represented at the level of individual neurons and neural populations in the avian brain. This comparative approach may provide insights into the general computational principles shared by biological and artificial systems for processing complex sequential communication signals.

Slot 16: Benjamin Hayden

Affiliation: Baylor College of Medicine
Department / Section: Neurosurgery
Job Title: Professor and McNair Scholar

Title: How can LLMs help us understand the brain

Abstract: Language requires us to maintain a structured representation of information that can be flexibly deployed in the service of both speaking and understanding. The question of how this information is represented and deployed is a central problem in neuroscience. The rise of large language models offers a useful set of hypotheses to test, and the development of single neuro recordings in human brains offers an opportunity to test these hypotheses. I will discuss recent findings about how language is represented in the human hippocampus – a region closely implicated in structured information representation – during listening and speaking. The hippocampus appears to instantiate a structured embedding space for language meaning, one that can be recovered even during anesthesia. In bilinguals, the geometry of the space is preserved, but rotated; the same is true when comparing speaking to listening. Moreover, the same space appears to be used for representing image semantics. Finally, techniques used for neuronal interpretability can help us to understand the contribution of single neurons to meaning representation.

Slot 17: Shinji Nishimoto

Affiliation: The University of Osaka / Center for Information and Neural Networks (CiNet) / National Institute of Information and Communications Technology (NICT)
Department / Section: Graduate School of Frontier Biosciences
Job Title: Professor / Executive Researcher

Title: Functional Organization of Perceptual and Cognitive Representations in the Human Brain across Naturalistic Experiences

Abstract: Human experience depends on the coordinated activity of distributed brain systems that transform sensory inputs into perceptual, semantic, emotional, and goal-directed representations. In this talk, I will present our efforts to characterize the functional organization of these representations using voxel-wise computational modeling across diverse naturalistic experiences.

First, I will discuss large-scale functional mapping based on 103 everyday perceptual and cognitive tasks. By systematically sampling functions ranging from sensation and action to memory, language, emotion, and decision-making, we identified a continuous and hierarchical organization of cognitive representations across the cortex. Subsequent analyses showed that this organization contains both broadly shared representational structures and reliable individual-specific patterns.

I will then extend this framework to more continuous and unconstrained experiences. Using hours of annotated movies and television dramas, we found that representations derived from large

language and vision–language models capture cortical processing across multiple levels, from concrete audiovisual content to abstract and long-timescale narrative information. Finally, analyses of spontaneous conversations revealed both shared neural representations for language production and comprehension and modality-specific organization across temporal scales.

Together, these studies suggest that diverse perceptual and cognitive functions are embedded within distributed representational spaces that remain systematically organized across structured tasks, continuous audiovisual experiences, and natural social communication.