# White Paper: Mathematics of Intelligences (IPAM Long Program, Fall 2024)

Montasir Abbas, Virginia Tech; Dalal Alharthi, University of Arizona; Daniela Beckelhymer, University of Minnesota;; Gregory Beylkin, University of Colorado Boulder; Polyphony Bruna, University of California, Merced (UC Merced); Erica Cartmill Indiana University Bloomington; Pattarawat Chormai, Technische Universität Berlin; Oliver Eberle, Technische Universität Berlin; James Evans, University of Chicago; Cynthia Flores, California State University, Channel Islands (CSU Channel Islands); Jacob Foster, Indiana University; Hamza Giaffar, University of California, San Diego (UCSD); Asvin Gothandaraman, Hebrew University; Mark Green, University of California, Los Angeles (UCLA); Linnéa Gyllingberg, Uppsala University; Olorundamilola Kazeem, University of California, Berkeley (UC Berkeley); Junsol Kim, University of Chicago; Luke Leckie, University of Bristol; Jiayi Li, University of California, Los Angeles (UCLA); Thomas McGee, University of California, Los Angeles (UCLA); Veronica Mierzejewski, Arizona State University; Ida Momennejad, Microsoft Research; Cody Moser, University of California, Merced (UC Merced); Omayra Ortega, Sonoma State University; Aishni Parab, University of California, Los Angeles (UCLA; Daniel Quigley, University of Wisconsin-Milwaukee; Paul Riechers, Beyond Institute for Theoretical Science; Danaja Rutar, University of Cambridge; Eric Schnell, London School of Economics and Political Science; Eder Sousa, The RAND Corporation; Peter Todd, Indiana University Bloomington; Joey Velez-Ginorio, University of Pennsylvania.

# 1. Executive Summary

The quest to understand intelligence is one of the great scientific endeavors, on par with quests to understand the origins of life or the foundations of the physical world. Several scientific communities have made significant progress on this quest. Relevant fields like animal cognition, cognitive science, collective intelligence, and artificial intelligence (AI)—as well as the social and behavioral sciences—have generated a wild variety of new experimental and observational data. They have also built mathematical and computational models of impressive sophistication and performance. Yet these communities remain largely disconnected; in no small part, this is because they lack a common framework and a shared (mathematical) language.

The IPAM Long Program on the Mathematics of Intelligences (MOI) aimed to bring these communities together with mathematicians to work toward the mathematical foundations necessary for transformational advances in our understanding of natural and artificial intelligences. This white paper was drafted at the culminating retreat of the Long Program and synthesizes the view of the field developed by its core participants. That said, it is not meant to be a comprehensive account of everything that happened at the Long Program. Likewise, the views expressed here do not necessarily reflect the views of IPAM or all the authors.

“Intelligence” is an ambiguous term. It can refer to an “intelligent *system*,” as when we speak of an “artificial intelligence.” It can also refer more broadly to a general *capacity*; roughly speaking, something that enables (intelligent) systems to solve problems more easily, whether they be organisms, collectives, or artificial agents. Although the former usage is returning to prominence in the age of Large Language Models (LLMs), we will generally refer to “intelligent systems” rather than “intelligences” to avoid confusion.

We also (attempt to) avoid the anthropocentric bias that makes humans the paradigm of intelligence. MOI explored intelligence as a multifaceted, multiscale phenomenon; it is for this reason that the Long Program was called the Mathematics of Intelligence**s**, with the final <s> embracing this multiplicity and variety.

Intelligent systems can take a wide variety of forms and formats: bees, birds, and bonobos; individual humans (of many ages and cultures); social networks and organizations; LLMs and interactive AI “societies.” They nevertheless display a set of shared capacities, including perception, learning, memory, reasoning, and creativity. In **Section 2**, we survey these capacities, illustrating them with examples from the MOI Workshops. We also indicate important *constraints* that shape how these capacities manifest and discuss a conceptual scheme (i.e., the Marr Levels) that informally organizes explanations of such capacities.

If there is to be a common framework that underlies the study of diverse natural and artificial intelligences, the language most promising for its formulation is mathematical. There have been a number of historical successes that provide a proof of concept: the Hodgkin-Huxley equations for individual neurons; imaging techniques for decoding functional Magnetic Resonance Imaging (fMRI); network analysis; Bayesian models of information processing in the brain; and (of course) artificial neural networks, from McCulloch-Pitts to contemporary deep learning architectures.

These are a handful of salient examples; **Section 3** surveys mathematical foundations and frameworks in much greater depth.

In **Section 4**, we examine biological intelligences including: the organization of structure in active matter and developing tissues; the computational capacities of neurons and other cells; perception, learning, and decision-making in human and non-human animals; and the evolutionary dynamics by which populations learn about their environments. Bayesian modeling, reinforcement learning, and dynamical systems are valuable frameworks for characterizing living systems, while physical models of cellular systems illuminate computation in matter. Many approaches focus on optimization; we also discuss homeostatic alternatives. We point to other promising avenues, including models of agent-environment feedback; porting concepts from AI into theories of biology; and the use of non-equilibrium statistical mechanics.

In **Section 5**, we explore diverse artificial intelligence architectures and the mathematical frameworks that might explain them, with an eye toward understanding AIs and aligning them with humans. We contrast deep learning and symbolic approaches, and compare Bayesian, agent-inspired, biologically-inspired, and physics-inspired architectures for understanding and constructing artificial intelligences. To probe these architectures, we explore frameworks as a prelude to building explainable AI and investigate methods to identify the underlying representation spaces and scaling relationships for these architectures.

In **Section 6**, we examine collective intelligence at multiple scales and degrees of complexity: the micro-scale (active matter and organismal structures); the meso-scale (population-level movement and coordination in flocks and ant colonies); and the macro-scale (human culture and broader social worlds). We focus on modeling frameworks from agent-based modeling, geometry, network science, and game theory. Fruitful future directions include more precise studies and models linking the cognitive strategies of agents to their influence on collectives; coordination between multiple LLMs and human agents; and drawing inspiration from models of coordination explicitly designed to understand communication in computer systems.

**Section 7** outlines attempts to recognize and assess intelligent capacities in diverse systems. We examine several key properties, including robustness, generalization, and the ability to balance learning strategies. We also highlight key challenges inherent in this approach: the difficulty of designing tasks that span different intelligent systems, and the difficulty of overcoming anthropocentric biases when designing metrics to assess intelligent behavior in unfamiliar contexts. We propose that a unification between frameworks from developmental and comparative psychology may be fruitfully employed in AI contexts and that higher-dimensional methods to identify intelligent behavior will be valuable to explore in greater depth.

We conclude in **Section 8** by summarizing some of the lessons learned from the Long Program, including promising directions in the mathematics of intelligences; major theoretical challenges; and the institutional and educational interventions required to make progress on this quest.

# 2. Theories, Capacities, and Examples of Intelligence(s)

Intelligent systems must solve a variety of problems. We can quantify their intelligence in two different ways. In a *narrow* sense, we can speak of how well a system solves a particular problem or class of problems. In a *broad* sense, we can quantify the variety of problems they solve, more or less well. Here we list several of the prominent capacities that intelligent systems use to solve problems. For each case, we describe the capacity; give examples from natural and artificial intelligences that were discussed during workshops and in working groups at IPAM; and point to theories and mathematical approaches that attempt to explain how the capacity works.

## 2.1 Some key cognitive capacities of intelligence

**Perception** is the ability of an intelligent system, natural or artificial, to acquire, process, and interpret information from its environment through different sensory modalities. For example, a dog recognizes its caregiver through their smell, and a self-driving car detects a pedestrian crossing the street using vision and lidar. Perception is involved in segmenting continuous input into meaningful units and objects. Statistical learning theory assumes segmentation either through chunking, where frequently co-occurring perceptual elements are grouped into cohesive units, or by tracking transitional probabilities between elements to identify boundaries. In both AI and biological vision models, individual neurons can become specialized for detecting specific features like edges, textures, or even high-level concepts like faces (T: Lu, W1: Yamins, T: Mehta; W3: Hillar). Information theory is a useful mathematical tool for understanding perception.

**Learning** is the ability to acquire new knowledge or behaviors by observing the environment or by adapting existing mental representations for a novel situation. A unicellular organism can synchronize its behavior to periodic environmental events, learning even without a nervous system (W1: Garnier). Another example is a recommendation system on a streaming service, which learns user preferences by analyzing viewing habits and suggests new content based on patterns in aggregated data. Learning theories include trial and error, reinforcement learning (W1: Momennejad), supervised and unsupervised learning, social learning through interaction and imitation, constructing and revising models or theories, and forming associations and recognizing statistical patterns. Learning efficiency measures how effectively a system acquires and generalizes knowledge, incorporating metrics like loss optimization, entropy and mutual information. Dynamical systems theory can model learning agents as following trajectories in a high dimensional conceptual space, for example in evolutionary models using random walks or in machine learning using gradient descent. (W3: Lelis; W4: Vogelstein)

**Concepts** are latent abstractions that help individuals condense, organize, categorize, and interpret perceptual content. Their compositional nature may be key to how humans and AI systems reason and learn (W3: Lewis, Tull). Concepts can be represented in implicit ways such as embedding vectors, or in highly structured forms like hierarchical structures or programs. Semantic embedding models like word2vec map words (and other linguistic units) to vectors based on context patterns, with more recent models using attention mechanisms to create more nuanced, context-dependent representations (W1: Foster). Representing concepts as programs

enables efficient search for useful hypotheses incorporating them. (W3: Lelis, Ellis, Chaudhari; T: Lupyan, T: Lu)

**Memory** is the capacity to encode, store, and retrieve information for use in prediction and guiding future behavior and decision-making. For example, paper wasps remember the faces of other wasps to enable them to navigate their social hierarchy (T: Cartmill). In human memory the hippocampus is involved in replay, pattern separation, and associative learning with long-term memories and spatial representations, and the prefrontal cortex with working memory and goal-directed tasks. Artificial agents implement memory through related architectures such as recurrent neural networks, LSTMs, and transformers. Computational models of memory formalize retrieval processes as searches through high-dimensional representations of learned information, optimized for efficiency and accuracy (W2: Krotov). In both biological and artificial systems, memory underpins functions such as reasoning, planning, and prediction (W1: Momennejad).

**Social intelligence and communication** govern how agents (human, animal, artificial) interact with others. For example, humans infer others' intentions from their actions, and use these inferences to multiple ends: to open a door for someone, to deceive someone, or to adjust their language to best communicate with a group (W2: Shafto; T: Lupyan). Scrub jays and chimpanzees both take into account what others can see when choosing where to forage for or hide food (T: Cartmill). Such inferences are equally important for artificial intelligence, where an ongoing challenge is developing agents that not only exhibit competence but also interact safely and ethically within a society (W1: Leibo; Kleiman-Weiner). One approach for modeling these interactions is through reinforcement learning (RL), which can be used to design systems to conform to particular goals (e.g. RL from human feedback) or to infer others' goals from how they behave (e.g. inverse RL).

**Search** is the active seeking of a desired resource that the agent does not currently possess, and does not know the location of. When a food-searching bear forages on berry bushes to find the most berries in a given amount of time, the bear must decide how long to exploit the berry bush it is currently in, and when to switch to exploring for a new berry bush after the first is depleted. The same kind of exploration/exploitation tradeoff may explain how a person searches through their memory for concepts occurring in clusters in high-dimensional concept space, when a curious child explores actions that can be performed with a new toy, and when groups seek novel solutions to problems (W1: Todd). The marginal value theorem prescribes the optimal point at which to switch from exploiting to exploring, and Lévy flight random walks can approximate such behavior. Artificial systems search for efficient paths through networks or graphs, such as seeking good routes through traffic-clogged streets. Optimal stopping and satisficing methods can determine when the best or a good-enough route has been found.

**Planning** is often operationalized as the capacity of an organism (e.g., rodent, human) to take a sequence of actions in an environment to achieve goals (e.g., food), where not all possible outcomes are visible from the starting point. It can be implemented as an internal search over possible trajectories and outcomes of actions. Computational frameworks such as RL formalize planning in terms of algorithms applied to learned representations in memory (W1: Momennejad)..

**Prediction** means anticipating future states, events, or sensory inputs based on current and past information. People perform predictive processing when understanding sentences: When we hear "The boy will eat the cake," our eyes move to edible objects before the word "cake" is spoken, showing real-time prediction of likely sentence completions. Similarly, predictive training lies at the core of modern AI systems including large language models that learn by attempting to predict masked words, next tokens, or future states based on context. Predictive coding theory proposes very generally that intelligent systems use generative models to actively predict incoming information rather than passively processing it. Computational mechanics offers a mathematical framework for intelligently balancing prediction accuracy with computational costs.

**Reasoning** involves the manipulation of concepts or representations to make judgments and decisions, update knowledge, and make predictions and plans about future states. For example, a crow might see a nut out of reach in a hole and select a stick of appropriate length to pry the nut out of the hole. In LLMs, explicitly inducing reasoning through methods like chain-of-thought prompts can improve their capabilities. *Deductive* reasoning is characterized by a top-down approach, drawing specific conclusions from general rules. *Inductive* reasoning takes a bottom-up approach, generalizing global rules from assessment of local information. *Abductive* reasoning infers the most likely explanation for a set of examples. Reasoning can be modeled as synthesizing programs, breaking down complex problems into smaller, manageable components (W3: Ellis), or as Bayesian inference, updating the likelihood of hypotheses based on new evidence (W3: Tenenbaum, Griffiths).

**Creativity**, the capacity to generate novel useful conceptions, is a hallmark of intelligence across natural and artificial systems. In individual human intelligence, creativity can emerge through analogical reasoning and conceptual blending, as when Einstein imagined riding alongside a light beam to develop special relativity. In collective intelligence, creativity often stems from the collision of distant ideas, as when scientific teams bridge disciplinary boundaries to forge breakthroughs via collective abduction, resolving surprises or problems from one field by theories, methods, or patterns from another (W1: Evans). In artificial intelligence, creativity manifests in systems that identify surprising but valuable combinations, like GPT-4 combining concepts to generate novel metaphors or DALL-E blending visual elements into novel artistic compositions (W3: Mehrotra). Creativity can be measured by geometric distances in learned semantic embeddings. Inductive and abductive theories show how creativity requires intelligence to generalize from novel cases or to connect previously disconnected representations while preserving their functional utility.

## 2.2 Work done at IPAM

Participants in the long program explored these capabilities as part of workshop lectures and in the working groups. Workshop 1 covered search, reasoning, problem-solving, perception, and creativity in individuals and groups of humans, AIs, animals, and cells. Workshop 2 focused on learning in artificial neural networks and brains. Workshop 3 delved into learning, concepts, reasoning, prediction, communication, and creativity in humans and AI, in both cooperative and adversarial settings. Workshop 4 covered foraging, planning, and reasoning in collectives from

chemical systems and insect swarms to computer networks.

## 2.3 Outlook: Levels & Constraints

How systems express these capacities, and how competently they use them, are profoundly affected by constraints limiting their access to key resources: computation, memory, and time (W3: Griffiths). Computational constraints privilege the efficiency of algorithms used to solve a problem (e.g., using a data-efficient learning algorithm or solving a search problem with a simple heuristic). Memory constraints affect the strategies that are viable (e.g., a scrub jay can remember the location of hundreds of cached nuts, while humans rely on external memory aids). And time constraints affect all aspects of problem solving: how quickly a decision must be made determines how much thinking can go into the decision, while how long an organism lives will affect how it trades off exploration for new knowledge versus exploitation of established knowledge. Different intelligent systems balance constraints in different ways. Living systems can use stored inductive biases (built-in over evolutionary time) that allow them to learn new skills quickly and efficiently, trading off memory for learning time and computation. Artificial agents can do something similar through the addition of memory modules.

While these constraints help to explain how a particular intelligent system manifests a capacity, we are still confronted with a multiplicity of explanations for any given capacity. Such explanations are not *necessarily* in conflict, because capacities can be described and studied at three different levels of analysis, as famously noted by the computational neuroscientist David Marr. At the *computational* level, we define the nature of the problem that the system is trying to solve, and the broad strategy of solution. At the *algorithmic/representational* level, we specify how the system represents information and what procedures (algorithms) it uses to transform these representations to solve the problem. Finally, at the *implementation* level, we describe how these algorithms and representations are physically realized in hardware, including biological tissue or computer circuits. Marr argued that these levels are relatively independent but complementary, with a complete understanding of any information processing system requiring analysis at all three levels. While this framework is widely used across cognitive science and AI, it lacks a formal mathematical account of how explanations at different Marr levels constrain and relate to each other; this is a key area needing development for theories of intelligence (Working Group: Rosetta Stone).

The next several sections of the white paper will explore mathematical theories of these intelligent capacities in detail (in both natural and artificial intelligences). Although preliminary accounts of resource constraints have emerged in recent years, we are nowhere near a complete theory that would connect the constraints, trade-offs, and history of a concrete intelligent system to the specific ways a capacity manifests in that system.

# 3. Mathematical Foundations for Understanding Intelligences

## 3.1 Introduction

This section surveys key mathematical paradigms for understanding models of intelligent systems, ranging from the standard linear algebra techniques to geometric and topological methods, probabilistic frameworks, and emerging theoretical constructs. By examining how information is represented, how learning converges to particular structures, and how high-dimensional, nonlinear problems can be managed, we lay a mathematical groundwork for integrative theories of intelligence.

Understanding intelligence, natural or artificial, requires novel theoretical frameworks and mathematical foundations. Genuinely new mathematical ideas will likely be needed to answer the grand questions of intelligence, but we are already learning much about intelligence from the perspective of existing mathematical fields. Many research endeavors strive to characterize, formalize, and unify principles that underlie learning, perception, inference, decision making, and so on. Understanding the mathematics of intelligence, therefore, involves a variety of mathematical techniques drawn from a wide array of mathematical subfields. A successful research project often draws from only one or a small number of these. Nature, however, is under no obligation to be constrained by our ingenuity; this is an invitation for innovation. It is our hope that this invitation inspires creativity and collaboration across mathematical domains, which are informed and remain informed by the domain sciences, for studying natural and artificial intelligences.

## 3.2 Work done at IPAM

The long program and all workshops touched on a broad range of mathematical fields, for which we provide a summary of sources of mathematical foundations (not a complete list) in the table below. A range of mathematical directions was explored by Working Groups on Algebraic Geometry for Machine Learning; Homeostatic and Ecological Intelligence (HEI); Emergence & Phase Transitions; Communication; Universal Representations; the Thermodynamics of Intelligences; and Rosetta Stone (an explicit search for mathematical connections between frameworks). Much more was covered in these working groups than we can discuss in this exposition. To give one example: the HEI working group recognized that living systems do not always present obvious optimization-based objective functions. Consequently, this group concentrated on Koopman operator methods to identify and predict cyclic patterns within data, thereby advancing our understanding of HEI-related phenomena.

**Survey of Mathematics for Understanding Intelligences (items arranged alphabetically)**

| Discipline/Domain | (Selected) underlying Theories and Techniques | Sample Applications |
|---|---|---|
| Algebraic Geometry | Polynomial optimization; algebraic varieties; resolution of singularity | Singular learning theory; scaling limits; optimization landscape |
| Analysis | Convolutions; transformations; regularity; approximation theory | CNNs; spectral analysis; approximation; filters in Fourier space; neural tangent kernel analysis |
| Category Theory | Functorial mappings; compositional structures; a uniform language for describing mathematical constructions | Unifying heterogeneous models (neuro-symbolic); language for expressing structure and relationships; composable description of institutions |
| Control Theory | Differential equations; transfer functions and transforms; structured matrices; dynamics and stability analysis; optimal control | State space models; robotics and adaptive systems; homeostatic models; computational neuroscience |
| Dynamical Systems | Stability analysis; bifurcation theory; Lyapunov functions; attractors; chaos theory; connections to Koopman Theory (DMD); nonlinear operators | Modeling temporal evolution of learning dynamics; neural and behavioral processes; various operators can be used to capture short- & long-term interactions |
| Game Theory | Nash equilibria; mean field games; evolutionary stable strategies (ESS) | Evolution of Learning; decision-making in multi-agent systems; reinforcement learning for artificial agents; economic and evolutionary dynamics |
| Geometry | Differential manifolds; metric spaces; curvature (negative-hyperbolic, positive-elliptical); dimensionality reduction; affine and projective geometries | Understanding learned feature manifolds/latent space in artificial neural networks; representations of semantic spaces |
| Graph Theory | Network properties (centrality, modularity); spectral analysis; evolution of graphs; graphical models; graph neural networks | Modeling neural connectivity and interactions of agents; analyzing knowledge graphs; social networks; semantic relationships; signal transmission |
| High Dimensional Linear Algebra & Functional Analysis | Johnson-Lindenstrauss lemma and consequences; sparsity; random matrix theory; spin glass models; Marchenko-Pastur law | Artificial neural networks; loss landscape; convergence analysis; scaling laws; operator approximations; dimension reduction; |

| | | |
|---|---|---|
| Optimization & Numerical Analysis | Gradient-based optimization; convex analysis; approximation methods; evolutionary algorithms | Training neural networks; tuning parameters efficiently; optimal transport |
| Partial Differential Equations | Harmonic analysis; nonlocal analysis; dispersive equations; mathematical physics | Wave propagation; elasticity; modeling temporal and spatial evolution of learning; pattern formation |
| Probabilistic Models & Statistics | Bayesian inference; MCMC; HMMs; variational methods; maximum likelihood; stochastic differential equations; statistical learning theory | Handling uncertainty; learning probabilistic models of cognition; stochastic gradient descent; randomized algorithms; artificial neural networks |
| Programming and Formal Language Theory, Logic | Type theory; lambda calculus; Operational semantics; denotational semantics; model theory | Concept learning; implementations; mechanistic interpretability; models of knowledge and reasoning; syntax/semantics interface |
| Topology | Persistent homology; topological data analysis | Uncovering the global structure of representations (e.g. data clusters) |

**Sample Sources of Mathematical Modeling Techniques**

| **Math Modeling Framework** | **Underlying Theories and Techniques** | **Sample Applications** |
|---|---|---|
| Agent-based Models | Self-propelled particle models; cellular automata; statistical physics | Collective intelligence; collective behaviour |
| Evolutionary Algorithms | Darwinian principles (selection, mutation, crossover); fitness landscapes; population-based search methods | Optimizing neural network architectures; evolving efficient robotics controllers; automated feature selection in machine learning |
| Reservoir Computing | Echo state networks; liquid state machines; recurrent neural networks; dynamical systems theory | Integrating multiscale temporal patterns; chaotic time series forecasting; signal denoising; classification; local computation; physical computation; neuromorphic/edge computing |

## 3.3 Outlook

The study of natural and artificial intelligences invites new formalisms that transcend existing frameworks, largely because the phenomena we seek to understand cannot be fully captured by existing mathematical tools. How do intelligences internally represent knowledge and concepts?

Which representations are well-adapted to finding what has been experienced? How do we model this mathematically? Do we have the right mathematics to do this? Are there universal representations that appear spontaneously? (Interestingly, there is some preliminary evidence for this last point, in the field of representation learning; e.g., W2: Reichers)

Difficult applied problems have a history of inspiring the development of important new mathematics. Likewise, having a well-formulated mathematical framework suggests the next useful experiments and simulations to conduct. Forging these new theoretical paths is hindered by a lack of common language across fields, suggesting the development of new venues for training the next generation of mathematicians and scientists studying artificial and natural intelligences. Addressing this requires a culture of mutual appreciation and collaborative efforts.

Several areas present fruitful avenues for future research. For example, the mathematician Gian-Carlo Rota noted decades ago that we lack an adequate formalization of relations like “as” that underpin core capacities like abstraction and analogy (e.g., seeing a piece of metal “as” a key). Ideas from category theory and dependent type theory may offer new avenues to conceptualize and formalize key facets of intelligence, though their roles are still largely exploratory (Opening Day: Foster; W3: Bradley, Lewis, Spivak; W4: Tan). While empirical scaling laws in AI settings have guided much of our understanding of how model size and data volume affect performance (many talks in Workshop 2), the precise interplay of model parameters, data complexity, and emergent properties remains elusive (W2: Hanin). Singular theory from algebraic geometry can potentially be leveraged to characterize this relationship (W2: Lin). Beyond artificial intelligence, the intelligent behavior of biological systems can also be modeled mathematically through techniques discussed at MOI, providing a broader context for how diverse mathematical frameworks might inform our understanding of intelligence (W1, W3, and W4 contained many examples along these lines). Going forward, we invite mathematicians across all domains to work and collaborate on challenging questions about natural and artificial intelligences, taking us closer to a unified *mathematics of intelligences*.

# 4. Models of Biological, Cognitive, & Ecological Intelligence

## 4.1 Introduction

Common notions of natural intelligence focus on cognition in organisms with nervous systems and brains. They identify intelligence in terms of cognitive capacities and strategies such as memory, reasoning, search (W1: Todd), and problem solving, among others. Mathematical models of individual behavioral, neural, and cognitive phenomena include: Bayesian approaches (W1: Lu, Allen, Gerstenberg; W3: Tenenbaum, Griffiths) and Reinforcement Learning (RL) (W1: Momennejad, McNamee) for modeling human and animal behavior, as well as dynamical systems models of embodied cognition (Tutorials: Bongard, W1: Garnier). Bayesian methods have extensions to program synthesis (W3: Ellis) and active inference (W4: Heins). Various mathematical frameworks also model neural recordings and fMRI using machine learning and mathematical modeling (W1: Mehta, Momennejad; W4: Couzin), as well as artificial neural networks (W3: Hillar, Barron). Additional frameworks include phase field modeling to understand cognition in slime molds (W1: Garnier), game theoretical approaches to decision making (W4:

Mann), self-propelled particle models to study collective animal behaviour (W1: Biro, W4: Couzin), and coupled oscillator theory to understand synchronization in fire flies (W4: Peleg).

We discussed possible ways to recast and mathematically model intelligence in terms of adaptive strategies that span two levels: evolutionary and organismal. The former concerns adaptive strategies at the scale of the species over evolutionary time; the latter concerns how organisms implement features of intelligence within a lifespan at individual and collective levels, e.g., learning and memory via cellular plasticity or problem solving by an ant colony or a group of scholars. We also discussed intelligence in organisms without brains or nervous systems (W3: Levin).

## 4.2 Mathematical models of cognition and brains

Mathematical and computational modeling have become pivotal in advancing our understanding of psychological, cognitive, and neural processes (e.g., mathematical psychology; computational neuroscience). For decades, this area has generated rigorous frameworks for normative, mechanistic, and predictive modeling of cognitive capacities and neural phenomena. In psychology and cognitive science, Bayesian models (T: Lu; W1: Gerstenberg, Kleiman-Weiner; W3: Tenenbaum, Griffiths) offer insights into probabilistic reasoning, perception, and decision-making, while reinforcement learning (RL; W1: Momennejad) elucidates reward processing and adaptive behavior as well as the structure of memory and learned representations in humans and animals. These frameworks extend to active inference, where organisms minimize prediction errors about their environment. Moreover, dynamical systems approaches, including phase space analyses and attractor models, have deepened our understanding of embodied cognition and neural coordination. Models derived from machine learning and artificial neural networks are used to model or analyze neuroimaging data across scales and species, from functional Magnetic Resonance Imaging (fMRI) to electrophysiology. Advanced approaches like phase field models and coupled oscillator theory attempt to bridge individual and group dynamics in biological and artificial systems.

Prominently, computational neuroscience and neuroAI (T: Mehta; W1: Momennejad; W2: Mehta; W3: Dasgupta, Barron) have revolutionized the study of brain function, bridging biology and artificial intelligence to model and analyze neural processes with precision. Computational neuroscience employs mathematical models and simulations to explore how neural systems and circuits encode, process, and transmit information. These include spiking neural networks mimicking real neuron dynamics, dynamical systems for network stability and oscillations, and energy-based frameworks describing neural interactions through optimization principles like free energy minimization. Machine learning, particularly deep learning, uncovers patterns in neural data, enabling researchers to map connectivity, decode representations, and predict behavior from neural activity.

NeuroAI expands these efforts by applying neuroscience insights to develop artificial neural networks and by using AI to explore biological intelligence. It avoids assuming a naive isomorphism between neural and computational systems. For instance, recurrent neural networks (RNNs) and transformer models simulate core functions like working memory, sequence prediction, and temporal processing, reflecting prefrontal cortex and hippocampal roles. NeuroAI

innovations focus on biologically inspired mechanisms such as hierarchical processing, sparsity, and neuromodulation to replicate the brain's efficiency and adaptability. Techniques like neural manifold analysis reveal how brains and AI systems simplify high-dimensional input to support decision-making and generalization. Drawing on evolutionary and adaptive properties like Hebbian learning, plasticity, and distributed processing, neuroAI continues to drive discoveries, including parallels between reinforcement learning in dopaminergic systems and gradient descent in artificial networks. These interdisciplinary approaches advance cognition research and guide the development of AI systems that embody principles of homeostasis, resilience, and ecological intelligence, pushing the boundaries of both fields.

## 4.3 Evolutionary and ecological adaptation

Evolutionary and ecological processes drive the emergence of innovation and biodiversity across diverse timescales. While there is no comprehensive formal framework for these interrelated phenomena, certain aspects of microevolution and ecology have yielded to mathematical analysis. Two established approaches stand out: population genetics and replicator dynamics, which model short-term adaptation (including inter-agent interactions); and evolutionary algorithms, which implement a range of evolutionary processes like selection, recombination, and mutation (W1: Bongard, W3: Forrest, O'Reilly). Other phenomena to capture at the evolutionary level include ecologically adaptive strategies such as mutualism, e.g., the Hadza hunter-gatherer community's collaboration with honey-guide birds to capture honey in a sustainable manner, and the especially interesting phenomenon niche construction, in which organisms modify their environment and thereby shape the selective pressures they experience (e.g., beavers building a dam or humans building cities). We regard the latter as an especially promising topic for innovative modeling and analysis.

## 4.4 Physics of living systems

Living systems can be described as physical systems, when viewed in terms of energy, information, and change. At the macro-scale, such approaches might attempt to identify energy flows within ecosystems and understand how these shape system-level resilience; or use techniques from statistical mechanics to describe and understand population-level changes in genes. On a meso-scale, organisms maintain balance between their environment and metabolism through self-maintenance and other capacities that enable living processes. On the smallest scale, with a focus on information and its replication, some researchers believe that physics-based models can describe the relationship between mutation, transmission, and error catastrophe rates, helping us understand the feasibility of replicators. Others consider physics inadequate at capturing all aspects of living systems, pointing to the emergence of autonomous laws or degrees of freedom that may obey distinct principles.

## 4.5 Can AI inform our understanding of biology?

There may exist deep correspondences between the ways that artificial and biological systems process information that invite mathematical exploration and characterization. Both biological and artificial systems, for example, instantiate a complex interplay of functions, with chemical

reactions serving as the base of biological systems and mathematical functions composing artificial models (W3: Forrest, W4: Menon). As functional ensembles, both systems must cope with high dimensional data and isolate task-relevant features. Perhaps, despite differences in substrate, there are generalizable laws that dictate how learning systems of sufficient complexity process information and decompose the world, from cells navigating through a vast transcription space to solve a novel problem (W3: Levin) to LLMs navigating through a vector space to predict the next token (many talks in W2). Probing features of artificial networks, such as the compression of high-dimensional data into more navigable low-dimensional spaces, the representation of features across layers (W1: Momennejad, W2: Eberle), compositionality, and generalization to out-of-distribution tasks (W2: Belkin, Radhakrishnan) can provide rich research questions for understanding the fundamentals of biological systems.

## 4.6 Work done at IPAM

Members formed overlapping working groups on homeostatic and ecological intelligence, thermodynamics, communication, and search, where they read and discussed the literature, invited external speakers, and discussed distinctions and similarities of adaptive strategies across evolutionary and organismal levels, as well as the relationship between models in physics to those employed in biology. Classical accounts often define intelligence in terms of maximizing short-term gains in an individualist, zero-sum manner. They assume a linear notion of time and a cumulative notion of rewards. While this account has been successful in developing contemporary artificial intelligence, it ignores a fundamental aspect of all living systems: homeostasis. Living organisms need to maintain dynamic homeostasis in long-term interactions with ecology as complex adaptive systems. The thermodynamics working group additionally discussed differing views of a possible relationship between physics and living processes and its caveats. They examined developments in nonequilibrium thermodynamics and dissipative structures for examining living systems at all levels (W4: Blair, Fakhri). The search group explored similarities and differences among approaches used by intelligent systems to seek and find the resources they need to maintain their ongoing existence (e.g., energy, information, social partners). The communication working group developed theoretical models of communication more appropriate for understanding non-human animals; formal semantic frameworks that contrast imperative and declarative communication; and frameworks from theoretical computer science that extend the classic one-way Shannon information transfer to allow for contextual modulation, feedback, inference, and common ground.

One participant proposed a paradigm shift from intelligence as reward maximization to “Ecological Intelligence,” which became the focus of a working group. The mathematical formalization of ecological intelligence focuses on homeostasis and adapting to replenishing cycles, not maximizing individual outcomes for short-term goals, nor satisficing (as in ecological rationality). A number of members worked on developing models of ecological intelligence rooted in homeostasis, evolutionary dynamics, and niche construction. One collaboration modeled these processes in cycles of depleting and replenishing resources in ecosystems and in far from equilibrium conditions. Another collaboration modeled local prediction via environmental entrainment in non-neural systems using metabolically-inspired allostatic learning rules to explain empirical observations in slime mold behavior. Lastly, IPAM workshops spurred discussion

around the collective intelligence of organisms such as locusts, fish, bees, and birds, while talks addressed the evolution of biological algorithms that process information, notably the recent advances in understanding the Drosophila olfactory system (W3: Dasgupta) and neural representations underlying decision-making in several organisms (W4: Couzin).

### 4.7 Outlook

A key future direction concerns mathematical frameworks for moving beyond simple optimization. What is the right mathematical framework for understanding intelligence in homeostatic systems? Mathematical models at the evolutionary scale often rely on hill-climbing on predefined static fitness landscapes (*à la* NK models), leaving a void for models that tether dynamic fitness to dynamic homeostasis within ecological niches. One core participant developed a book proposal on Ecological Intelligence, which they will expand into a book. The aforementioned ongoing collaborations on modeling and experimentation fostered during the IPAM Long Program begin to fill the void around evolution and homeostatis in dynamic and continually constructed fitness landscapes.

# 5. Theory and Architecture of Artificial Intelligence (AI)

## 5.1 Introduction

Modern AI, emerging from connectionist principles, has been shaped by advances in mathematical theories, parallel computing architectures, and data availability. While deep learning has achieved remarkable successes through gradient-based optimization and high-dimensional vector operations, its complete mathematical framework remains under development. Current approaches focus on function representation and data processing methods, with practical implementations bounded by computational constraints. However, our limited understanding of contemporary AI's internal representational mechanisms continues to raise important questions about safety and reliability in real-world applications, while encouraging the development of alternative approaches to AI. The following sections aim to summarize current and emerging AI architectures and directions, their theoretical foundations, and practical implications, while highlighting recent advancements and ongoing research presented during the program.

## 5.2 Architectures

The development of AI architectures has evolved through a dynamic interplay of bottom-up and top-down approaches, each reflecting distinct principles of intelligence. Since the early 2010s, bottom-up architectures characterized by large-scale deep learning models leverage self-supervised training on vast unstructured datasets and reinforcement learning (RL) to produce foundation models capable of generalizing across diverse tasks. The success of these architectures is in part thanks to scalability, and the fact that they can be further improved with fine-tuning and continual learning (W2: Smith). While deep learning architectures dominate modern AI, their reliance on opaque inner workings and significant computational resources

highlights critical challenges. In contrast, top-down architectures integrate domain-specific knowledge from fields like neuroscience, physics, and cognitive science. These architectures emphasize interpretability, efficiency, and robustness, but often struggle with scalability. Hybrid neuro-symbolic architectures attempt to bridge these paradigms, combining neural networks with symbolic reasoning to enhance scalability, structure and transparency (W3: Chaudhari, Ellis, O'Reilly, Gulwani). Bayesian architectures, which model uncertainty and update beliefs with limited data, provide an alternative framework for reasoning under uncertainty but face challenges in handling large-scale tasks (W1: Allen; W3: Griffiths, Tenenbaum). Biologically-inspired architectures draw from principles like sensory invariances and evolutionary dynamics, while physics-informed architectures, such as PINNs, merge data-driven learning with fundamental physical laws to address complex scientific problems. Together, these evolving frameworks for designing AI architectures reflect a broader rethinking of AI's foundations and their potential applications.

## 5.3 Theory and Understanding of AI

### 5.3.1 Mathematics of DL

Modern deep learning (DL) models differ from classical machine learning models in that they commonly operate in an over-parameterized regime, where the number of parameters exceeds the amount of training data. In this setting, the optimization landscape is highly non-convex and may admit multiple interpolating solutions—each perfectly fitting the training set. Despite this complexity, we expect DL models to also perform well on previously unseen data that share the "expected features." Empirically, different interpolating solutions can exhibit vastly different generalization abilities. Nonetheless, gradient-based optimization methods often find solutions that generalize surprisingly well. This phenomenon is known as the "implicit bias" of gradient descent, and building a rigorous theoretical understanding of it remains an active area of research (W2: Tsivilis). Other central open questions include providing convergence guarantees for optimization algorithms, characterizing benign overfitting (W2: Murray), and exploring related foundational issues.

Empirical results from large-scale training further illustrate the complexity of DL phenomena and have inspired the development of Scaling Laws (W2: Hanin, Dohmatob, Pehlevan, Roy, Bordelon, Yang, Vankadara). These laws indicate that, given sufficient model parameters and data, models will reach performances and emergent abilities that are not achievable in smaller models. However, this approach often demands enormous computational and energy resources, limiting the accessibility and sustainability of such systems. Moreover, it challenges existing theoretical frameworks. Drawing inspiration from natural intelligence—where intelligence emerges and thrives under tight resource constraints—could lead to more sustainable directions for advancing and understanding AI.

### 5.3.2 Interpretability and Biases

Despite AI models achieving superhuman performance in many tasks, understanding, interpreting, and controlling them remains challenging, raising concerns about safety and the amplification of biases in training data. While explainable AI offers insights into model predictions, fully understanding their internal mechanisms and data interactions remains a key technical hurdle (W2: Belkin, Eberle, Riechers, Radhakrishnan; W3 Tull, Gulwani). Moreover, AI systems can perpetuate and amplify existing societal biases through various mechanisms (W1: Needell). These biases stem from multiple sources: unrepresentative training data; historical discrimination embedded in datasets; lack of diversity among AI developers, researchers, and mathematicians more broadly (W1: Ortega); and feedback loops that reinforce existing prejudices. While some argue AI could help overcome human cognitive biases through data-driven decisions, the evidence suggests that without careful oversight and diverse input in development, AI systems risk automating and exacerbating discrimination across sectors.

### 5.3.3 Alignment

Alignment in cognitive science explores how individuals develop mutual understanding and organize their concepts and perceptions of the world. Unique experiences shape assumptions during communication, requiring individuals to negotiate new systems “on the fly” for shared tasks or problem-solving. In AI, by contrast, efforts focus on aligning neural representations and behaviors with human values, though defining these goals consistently is challenging. Techniques like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) improve language model usability but often result in superficial and fragile alignment. Achieving robust mathematical guarantees for alignment and safety remains a major challenge.

## 5.4 Work done at IPAM

Workshops 1 and 2 showcased advancements in generative AI models (W1: Evans, Leibo, Momennajad; W3: Riechers; Chaudhuri); novel methodologies based on insights from theoretical ML (W1: Foster); dense associative memory (W2: Krotov); kernel-based interpretations of neural networks; grokking (W2: Belkin); and interpretability (W2: Radhakrishnan, Eberle). Other topics included neuroscience-inspired generative AI architectures, diffusion models mimicking human behavior, and multi-agent systems using deep RL and LLMs (W1: Momennejad). Physics-inspired representations examined neural activation geometry (W2: Riechers). Workshop 1 and Workshop 3 also explored neuro-symbolic methods like symbolic regression, probabilistic programming, and program synthesis for inverse graphics (W3: Yildirim), physical cognition (W1: Allen), and other problems.

## 5.5 Outlook

The ArchEval working group focused on analyzing latent representations and attention in LLAMA 3.2 (7B) using neuroscience-inspired methods to identify an emergent search algorithm. Their work included tasks like explaining map or graph structures; goal-directed navigation; and

analyzing behavior, representations, and algorithms for optimal pathfinding. Several additional emergent teams have formed to work on 1) aspects of understanding the representation of AI; 2) novel AI architectures for a wide range of individual and collective intelligence tasks; and 3) state space models with a new algorithm for implementation developed during the program.

# 6. Collective and Multi-Agent Systems

## 6.1 Introduction

Many intelligent systems are characterized by a nested organization of agent collectives: collectives of cells form tissues and organs; arrangements of organs form individual organisms; individual organisms can further form social groups (W3: Levin). Acting as multi-agent systems, the inherent properties and actions of participant agents can give rise to an emergent **Collective Intelligence**. Here, we adopt the definition of collective intelligence as the "shared intelligence that emerges from the collaboration, collective efforts, and competition of many individuals." In essence, collective intelligence can endow multi-agent systems with capacities that were unavailable or limited in their constituent agents. Human organizations, for example, can successfully manage the production and distribution of products over global networks at a scale that would not be possible for a single individual. More recently, multi-agent LLM systems have demonstrated enhanced performance on cognitive tests as compared to single LLMs (W1: Momennejad, W1: Evans).

## 6.2 Biological collectives

Collective behavior and intelligence are expressed at varying scales and degrees of complexity. At the micro-scale, geometric and physical principles govern the self-assembly of molecules into their functional structures, while the active processes associated with life combine with physics to give rise to non-equilibrium matter that exhibits adaptive and emergent material properties. For example, the ability of viral capsid proteins to self-assemble is given by their two-dimensional geometry (W4: Menon), whilst starfish embryos aggregate into "chiral crystals" that comprise thousands of spinning organisms that break the symmetries inherent in non-living matter (W4: Fakhri). In systems such as these, geometry and active matter physics provide powerful tools to explore the individual-level mechanisms that give rise to collective dynamics.

At larger organismal scales, behaviours such as schooling in fish (W4) and bridge-building in ants (W1: Garnier) are visually stunning and adaptive expressions of collective intelligence. Yet, these behaviors arise from relatively simple individual-level properties. For instance, schooling and flocking equip bird and fish collectives with an expanded sensory range to globally detect and respond to predators. Yet, physics-based classical schooling and flocking models consider agents as self-propelled particles with local zones of attraction, repulsion, and orientation (as in the Couzin model; see also the Vicsek model). Contemporary models of collective movement more closely consider neural dynamics by explicitly modeling the torus-like arrangements of neurons into ring attractors. Such ring attractor models reproduce collective movement while considering only the angular separation between spatial loci or conspecifics (W4: Couzin).

At an increased level of social complexity, inter-individual interactions and task specialization (division of labour) give rise to complex social and spatial structures (W1: Velez). This is typified by the societies of social insects and humans. In these societies, vast social and spatial networks allow the distribution of resources and information across thousands of individuals (W4: LeBoeuf) whose interactions and activity are shaped by complex institutions (W4: Tan). The topology of multi-agent social networks is known to shape task performance, innovation, and memory (W1: Evans, Momennejad). In human societies in particular, the transmission of social information enables the distribution and development of culture.

## 6.3 Culture

Culture, defined as “socially transmitted information capable of affecting individuals’ behaviour”, is found in many intelligent species. Knowledge, beliefs, practices, and behavioral solutions can be directly transmitted from one member of a species to another; in humans, this often relies on cognitively sophisticated cooperative communication (W1: Hawkins; W2: Shafto) and a range of cognitive tools (W1: Fan) and cultural artifacts (W1: Foster). In the context of collective intelligence, culture enables the offloading of cognitive tasks in response to system-emerging properties and adopted changes. Instead of every community member being required to know all that is necessary to function and excel, knowledge and abilities can be divided up and reshared culturally. In certain cases, this broader cultural capacity leads to the emergence of cumulative culture, whereby cultural innovations that would be impossible for any one individual to develop can instead be developed by the collective, possibly over many generations (W1: Velez). Although debate on the topic remains, cumulative culture appears to be one of the unique capabilities of humans, and perhaps explains the emergence of human intelligence as a distinctively powerful form of biological intelligence.

Culture can be studied using a number of quantitative methods. One such framework is cultural evolution, which takes inspiration from biological evolution and uses many of its methods to study the manner in which culture changes and is transmitted. Such a framework has provided valuable mathematical rigor to the study of culture and has led to many insights into its role in intelligence. Other methods for studying culture include network approaches to understand the dynamics of social interaction and cultural transmission (W1: Evans); machine learning methods to represent and analyze complex cultural structures like narratives (W1: Foster); generative AI methods for modeling complex social interactions (W1: Leibo); and game theoretic approaches, in which cultural solutions emerge as strategies in different competitive and cooperative scenarios (W1: Kleiman-Weiner). Advances in AI have enabled the study of cultural data at a large scale, while potentially redefining culture itself.

## 6.4 Human-AI and AI-AI interactions

The capability of humans to create, interpret, and share (cumulative) cultures has been considered a crucial mechanism that enables human collectives to flourish. In the last few decades, we have observed machines, especially AI and LLM agents, emerge as cultural agents that generate and transmit cultures alongside human agents. A growing body of research suggests that LLMs can effectively learn and reproduce various cultural patterns, from linguistic

styles and social norms to complex behavioral patterns and ideas observed in human societies. With their sophisticated abilities to synthesize and generate novel cultural ideas, they are becoming both the medium in which culture transmits (e.g., recommender algorithms) as well as a generator of novel cultural ideas and products (e.g., generative AI that creates visual arts).

With the ability to interact and cooperate with other agents autonomously, “AI societies” can manifest in various forms, from collaborative problem-solving systems to interactive storytelling environments. Multiple AI agents can work together to create solutions (e.g., “storytelling” agent, “software developer” agent, and “designer” agents collaborate to build a game) or develop cultural narratives, such as scenarios of movies. AI agents often have different capabilities and perspectives in solving a particular problem due to different training data or model architectures (T: Page), which enables the combination of AI agents’ outputs to perform better than a single AI agent. As AI agents increasingly participate in multi-agent systems alongside humans and imitating humans in some instances, they create complex human-AI networks, where human and AI agents collaboratively solve problems and generate new cultures (W1: Evans; W4: Abbas and Alharthi).

## 6.5 Work done at IPAM

The themes discussed here arose primarily in Workshop 4: Modeling Multi-Scale Collective Intelligences, where participants discussed collective dynamics at length, ranging from particle self-assembly to social organization. Workshop 1: Analyzing High-dimensional Traces of Intelligent Behavior also presented innovations in measuring and modeling collective behaviors at many scales and degrees of complexity (e.g., Biro, Evans, Fan, Foster, Hawkins, Leibo, Kleinman-Weiner, Velez). The Cultural Dynamics WG explored the interplay of social structure, individual preferences, and cultural evolution with agent-based simulation. This working group also explored the role of cultural and linguistic backgrounds in shaping the social networks of LLMs. The Homeostatic and Ecological intelligence WG explored the emergence of intelligence in ecological settings, comprising multiple agents. The ArchEval group investigated the cognitive performance of different multi-agent LLM topologies. The Thermodynamics group explored physics models for collectives. Finally, the Communication group investigated novel frameworks for modeling inter-agent signalling and communication.

## 6.6 Outlook

Many intelligent species live in communities which develop collective and cultural intelligences. As we continue to develop tools for studying these collectives, we better understand how it is that information or knowledge is transmitted and shared across individuals. Human cultural dynamics exhibit complex evolutionary patterns, requiring mathematical frameworks to study them. As AI agents become increasingly integrated into our daily lives, human behavior and capabilities will adapt to function within a collectively intelligent multi-agent system. While we will gain new abilities, we may also lose some. To navigate these changes effectively, it is essential to develop a vision for understanding these technologies and setting priorities for research and development. For this, a wide range of tools from mathematics, physics, and complexity science—as well as the social and cultural sciences—can be applied towards the modeling and analysis of systems

at multiple scales. At increased scales of complexity, computational agent-based models, AI, and data science can offer myriad approaches for the study of cultural and sociological phenomena; whilst multi-agent LLMs are a promising new system for experimenting on complex social dynamics at scale.

# 7. Recognizing and Assessing Intelligences

## 7.1 Introduction

The fundamental challenge of studying intelligence lies not just in defining it, but in developing rigorous frameworks to detect its presence and evaluate its degree across diverse manifestations. This section explores multifaceted approaches to recognizing and assessing intelligence across biological, collective, and artificial systems, as well as their combinations. We examine how different theoretical traditions, from neuroscience to information theory, have shaped our evaluation methods while considering key dimensions of intelligence, including robustness, adaptivity, and communication. Special attention is given to the methodological challenges of comparing intelligences across different substrates, timescales, and task settings and also to the inherent biases present in historical and current evaluation frameworks that color traditional understandings of intelligence. By synthesizing insights from mechanistic interpretability, empirical observation, and cross-domain benchmarks, we aim to advance our ability to systematically assess and compare diverse forms of intelligence in diverse intelligent systems.

## 7.2 What to evaluate?

The quest to recognize and assess intelligence spans an expansive spectrum of simple and complex entities, from individual organisms to complex multi-agent systems. These topics structured the four workshops, from W1 (Analyzing High-dimensional Traces of Intelligent Behavior) and W2 (Theory and Practice of Deep Learning) to W3 (Naturalistic Approaches to Artificial Intelligence) and W4 (Modeling Multi-Scale Collective Intelligences); each workshop involved evaluations of intelligence across a range of intelligent systems. While human individuals and collectives have traditionally served as archetypal reference points for intelligence assessment (e.g., many talks in W1; W3: Griffith, Tenenbaum, Sadrzadeh), our understanding now encompasses the diverse cognitive capabilities of non-human animals and other organisms (Tutorial: Cartmill; many talks in W3); the emergent intelligence of collective and ecological systems (talks in W1 and W4); and the rapidly evolving landscape of artificial intelligence, especially generative AI systems (W2), which present novel challenges and opportunities for evaluation. Additionally, the growing integration of human and artificial intelligence is now creating hybrid ensembles that may demand their own frameworks for assessment, including considerations of human-centered AI design, interaction dynamics, alignment between human and machine cognition, and the effective integration of complementary capabilities for collective intelligence.

## 7.3 Traditions and Inspirations of Evaluation

Theoretical foundations for evaluating intelligence draw from multiple scientific and mathematical traditions, each offering unique insights, intelligence targets, and methodologies. Neuroscience provides mechanistic understanding of biological intelligence through neural architecture and dynamics, while cognitive science contributes frameworks for understanding mental processes and representations. Social sciences, including psychology, sociology, and anthropology, illuminate how intelligence manifests and evolves within cultural contexts and constructs protocols that form the basis for collective (human/AI) intelligence. Computer science offers operational definitions of intelligence, exemplified by the Turing Test's behavioral approach to evaluating human-like cognition. Comparative cognition research has developed sophisticated transfer tests and task batteries to assess intelligence across species, revealing commonalities and divergences in cognitive capabilities. Mathematical approaches provide formal frameworks for understanding intelligent inference and decision-making, including: Bayesian models of optimal rationality; reinforcement learning (RL) models; and information theoretic models that provide bounds for optimal information processing and learning capacity. Different "intelligent" systems may be differentially evaluated and compared using each of these formal frameworks (and more).

## 7.4 Dimensions of Intelligence to Evaluate

Recognizing and assessing intelligence requires understanding how systems navigate complex cognitive challenges across multiple dimensions. The cognitive "problem space" is shaped by temporal constraints that affect windows for decision-making; computational limitations that bound processing capacity; and communication barriers that limit or foreclose information exchange. When evaluating (potentially) intelligent systems, we must examine their organizational characteristics along several key dimensions. Of particular interest are: 1) system robustness (capacity of the system to perform despite shifts in the environment; is it resilient against adversarial attacks and internal component drift?); 2) Generalization beyond excelling in specialized tasks (truly intelligent systems demonstrate flexibility in resource allocation, successfully transferring knowledge across domains and generalizing from limited examples to novel situations); 3) Learning to learn (ability to strike a balance between exploration and exploitation, knowing when to leverage existing knowledge versus when to seek new information and deploy or test new approaches).

## 7.5 Challenges in comparison

Assessment of intelligence relies on comparisons. These can be within or between agents. The dimensions of comparison bring their own unique challenges to the problem of accurate and equitable assessment. The same individual or program can be compared across time, tasks, or contexts. Performance can also be compared across individuals, groups, species, or platforms. Designing equivalent tasks to assess intelligence does not always mean using identical tasks: a box-opening task used to assess physical cognition in primates relies heavily on manual dexterity

and could not be given to a species with a different body plan, like dolphins. In evaluating biological intelligences, careful consideration of natural behavior and ecology is critical in designing tasks to capture the full ability of a species. In evaluating AI, careful consideration of the training data and protocols can similarly inform task design and interpretation of task performance.

The act of comparison carries with it the implicit adoption of a baseline that guides or shapes the assessment. We humans tend to ground our search for intelligence in the things that we do well. Humans have big (centralized) brains and long developmental periods; we rely on vision and are manually dexterous; we make tools and easily manipulate symbols. In assessing AI systems and non-human animals, particular aspects of cognition are privileged if they are areas where humans excel (e.g., projection, causal inference, analogy) and are downplayed when they are not (e.g., spatial memory, rote memorization, olfactory perception). Historically, humans have been treated as a homogenous group, but further research has shown that an overreliance on western participants has distorted our understanding of human intelligence (the so-called "WEIRD" problem, indicating the non-representative nature of the **W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic subjects easily available to Western researchers). Researchers therefore tend to privilege the performance of *particular* human groups in the tasks that we most closely identify as markers of human intelligence. Making these human- and culture-centered biases visible can help broaden and diversify our search for and evaluation of intelligent capacities in non-human systems. In particular, considering the degree to which we intuitively privilege language as a hallmark of intelligence can help us design better comparisons between LLMs, Multi-modal models and non-linguistic systems.

## 7.6 Evaluating Intelligence: from Performance Metrics to Mechanisms

To evaluate intelligent abilities, we can distinguish between (i) top-down strategies that relate input-output evaluations of systems (can a system reliably produce a desired output given a specific input?) as compared to (ii) bottom-up evaluations that focus on the mechanistic processes implemented by a system. Evaluation of nominal task performance metrics such as accuracy provides one general but limited framework for comparing systems. Given the recent success of generative AI models on a variety of tasks, researchers have increasingly explored comparisons between their abilities and those of humans. Focusing solely on performance scores can be misleading, however; AI models are prone to learning seemingly accurate strategies that rely on undesired shortcuts. Understanding the underlying computational strategies and representations provides an important complementary direction to evaluate intelligent systems and verify sound task performance, as opposed to the traditional reliance on surface-level correlations in input-output performance. Given that many systems operate as black-boxes, methodologies to gain mechanistic insights into such systems are needed, ranging from interpretability approaches in AI (e.g., distillation) to reconstructing computational mechanisms of human and natural systems. Animal cognition and developmental researchers have developed a range of strategies for probing mechanisms in otherwise black-box systems (non-human animals or pre-verbal human infants).

### 7.7 Work done at IPAM

Participants brought to IPAM their experience studying a (truly) vast range of diverse systems and hailed from a wide variety of intellectual traditions and approaches. The ArchEval working group focused on the evaluation of cognitive capacities and algorithms in LLMs, multi-agent models of LLMs with different connectivity among components, and experiments that probed learning and memory in slime molds. The Rosetta Stone working group actively sought to reconcile diverse measurement theories and approaches to evaluate intelligence in different systems, with an eye to differentiating not a singular axis of intelligence but a range of consistently important axes and mathematical correspondences. All four workshops addressed intelligence evaluations, while working groups tackled questions in this space ranging from identifying high-dimensional traces of intelligence to evaluating search and planning in LLMs. Moreover, participants actively sought to apply mathematical concepts of intrinsic (necessary) dimensionality, linearity, and transport to existing evaluations in search of greater similarities.

### 7.8 Outlook

Evaluations are a booming area in the context of Artificial Intelligence systems. AI evaluation borrows from the deep scholarship on evaluation of human and non-human animal intelligences, though much more robust cross-fertilization is needed. In particular, the cultures in developmental and comparative psychology of systematically searching for alternative explanations for behavior should be widely adopted in AI. There is a capacious intellectual demand for new evaluations and evaluation methodologies as AI systems continue to evolve; by extension, there is a similar demand for new modes of evaluating the human, biological, and mixed systems to which AI systems are most intensively compared. Many participants in the long program will continue to collaborate and publish actively on these topics.

## 8. Conclusion: Unifying Themes, Meta-Questions, Philosophical Perspectives, and Interdisciplinary Implications

We learned several critical lessons from the Mathematics of Intelligences Long Program: about the variety of mathematical ideas that can be used; the substantial theoretical challenges faced by the development of this new area of investigation; and some of the institutional and educational challenges raised by this radically interdisciplinary approach.

Intelligent systems arise in many different contexts. It is therefore unsurprising that the useful mathematical methods and frameworks for studying these intelligences are similarly broad. Algebraic geometry, category theory, control theory, dynamical systems, programming theory, and topology are among the many fruitful mathematical approaches for understanding intelligent systems. This breadth of methods not only reflects the diverse subject matter to which these

methods are applied; it also reflects the variety of *levels* on which such systems can be analyzed. Methods used to understand and predict the *behavior* of cognitive agents are not necessarily the same as those used to understand their *generating mechanisms*. Compounding this issue, generating mechanisms can further be explored by examining their high level goals and strategies; the algorithms and representations that instantiate those strategies; and their implementation in specific "hardware" (whether biological or artificial).

Why should we expect this program to hang together scientifically? One might legitimately ask: To what extent do ideas helpful in understanding one intelligence translate to another, quite different, intelligence? There are a number of examples where this successful translation has indeed occurred, and during the Long Program we identified many formal and theoretical intersections between cognitive, biological, and AI models. As may be discerned from the sections above, we were able to find new instances of such concordances and have an informed optimism about connecting frameworks for studying diverse intelligences. This is not at all to say that there is a unitary framework which will work uniformly across all intelligences; instead, we assert that there is a great deal to be learned when we do not study intelligences in isolation.

How should we deal with this variety of frameworks and formalisms? Marr's levels remain the traditional scheme for organizing diverse explanations of intelligent systems; indeed, this has become a critical shorthand across the cognitive and neural sciences, and increasingly in AI. But it remains entirely informal. It would be immensely valuable to formalize the implicit notion that explanations on different levels *constrain* one another. For example, a Bayesian account (at the computational theory level) might be *implemented* by a particular approach to approximate Bayesian computation (at the algorithmic level) like MCMC, which may or may not be *implementable* on realistic neural hardware. Understanding such "implementation" relations emerged as a critical item of discussion for one of the working groups (which sought to explore a "Rosetta Stone" that would link different theories of intelligent systems together). Likewise, developing a deep formal understanding of the process of "translating" a framework from one domain to another–and grasping what formal properties are lost and retained in such a translation–remains a major mathematical and technical problem.

Despite the breadth of this program, it touched only lightly and informally on some of the social and ethical questions raised by the explosion of research on intelligent systems. These include pressing questions about the safety, accountability, bias, and ethical ramifications of artificial intelligence; the profound environmental impact of AI technologies (at least in their current compute- and data-hungry manifestation); the impact of AI on work, labor markets, and education; and issues of animal welfare and conservation raised by growing appreciation for the intelligence of non-human animals. Each of these topics represents a fruitful area for mathematical and computational scientists to engage with domain scientists, philosophers, social scientists, and policy researchers.

Another limitation of the program is its heavy reliance on the traditional formulation of intelligence as "problem solving." While critiques from homeostatic and ecological perspectives somewhat addressed this limitation, much is left to be done. Promising directions include: formalizing the capacity of intelligent systems to determine their own goals and set their own problems; better

understanding the way real-world environments are “read” by intelligent systems as sites that demand (or afford) particular activities; creating a mathematical language for talking about key capacities like creativity, imagination, or aesthetic appreciation; and much more.

This program was successful in assembling an interdisciplinary community of extraordinary breadth. Mathematicians, computer scientists, specialists in animal cognition, artists, neuroscientists, philosophers, bioscientists studying the origin of life, physicists, psychologists, cognitive scientists, and sociologists all worked together fruitfully. They did not self-segregate, but instead formed new and unexpected collaborations. Buoyed by this very positive outcome, we would suggest that collectively the organizers and participants learned a lot about how to put a project like this together and that this program in an emerging interdisciplinary field provides a model and a proof of concept for future projects of this depth and scope. At the same time, it also underlines that scientists from diverse backgrounds do not automatically work together effectively. Interdisciplinary collaboration is hard work, and further investment in training current and emerging scholars in interdisciplinary collaboration, as well as shared languages, concepts, and tools, would make a huge difference in the growth of this emerging field.