Integrative, Embodied Agents to Reverse-Engineer Natural Intelligence

Aran Nayebi

February 5, 2024

The overarching question that we want to work towards is: **What are the principles of autonomous, embodied intelligence that enable animals to transform continuous sensory inputs into meaningful physical actions, that can be abstracted into machines?** We operationalize this goal by reverse-engineering the algorithmic principles of animal cognition, and use what we find to empower more autonomous and adaptive robots. The following 9 pages form the introduction and motivation (constituting §0) to a longer lab plan that includes a detailed project list for lab members. I'm including the table of contents of the full document to give a sense of the topics we are interested in[1]. If this type of work interests you, email me to learn more about project specifics, or to collaborate!

Contents

0 Background

Why study this question? Evolution, spanning the phylogenetic tree from nematodes to fish, rodents, and primates, has consistently achieved what we have not yet been able to build – namely, embodied agents capable of flexibly and robustly interacting with the physical world to ensure their survival. This sensorimotor loop is the foundation of intelligence that's shared across species, and upon which our more abstract reasoning capabilities (including language) rest. However, engineering this capability has been a major computational challenge in artificial intelligence (AI), especially considering that it has been a long-held (yet unachieved) goal to build general-purpose robots. Despite algorithmic and dataset scale advances that enable effective representation learning [18], current AI struggle to understand the

[1]For either a short or long form (includes past work) video overview, see here or here, respectively.

world in physically-grounded, common-sense terms [34, 70]. Nor is existing AI as autonomous, or as able to learn[2], adapt[3], and act[4] in novel situations as animals do.

Central to this capability in animals is the structured interaction of neural circuits in their brain to bring forth complex behaviors (forming a "cognitive architecture"). Traditionally, neuroscientists have often focused their studies on individual brain areas, leading to a rich and detailed understanding of these specific regions. However, this focus has left understudied how these various brain areas collaborate to enable animals to build a model of the world and use it to execute actions within their environment. (Not to mention we do not truly understand any individual brain area without *also* understanding how it works with other brain areas[5].) This question has been particularly challenging to address in the past, as neural recording technologies were limited to capturing only small portions of a neural population in a single session. The development of large-scale recording techniques [149, 80], which now allow for simultaneous recordings across multiple brain areas in behaving animals, offer an unprecedented opportunity to reverse-engineer this phenomenon. Coupled with the advent of task-optimized artificial neural network models that yield accurate accounts of individual brain areas[6], we now have the potential to start to understand how these areas collectively work together.

To understand this interaction, we aim to gain a more comprehensive understanding of what constitutes intelligence by reverse-engineering it across multiple species. In other words, the best way to understand the brain is to try to *build* one. Specifically, we will reverse-engineer the foundational sensorimotor loop conserved across species through building integrative, embodied computational models, or "agents". Historically, there have been many agent proposals (e.g. ranging from the early days of cybernetics [170] and symbolic AI/robotics [105, 106, 123, 21], to recent ones in the language of modern deep learning [84, 153]). But none of them are specific enough to be built, let alone be predictive of neural activity (they're "not even wrong!"[7]). In other words, they aren't *testable* scientific theories of intelligence, so much as they are *frameworks* based on a useful collation of high-level takeaways from cognitive science, AI, and neuroscience. The hard part, and what forms the core of our research focus, is figuring out the specifics of the agent architecture as we build them. Therefore, our aim of building an embodied agent (cf. Figure 1), one for each species under study, forms the superstructure from which we generate scientific questions of interest. To get at this, I consider five main components of the agent that we will make concrete by specifying suitable tasks and architectures for, and which many species share:

[2]Especially learning a model of the world in an online, sample- and energy-efficient way, and updating it without catastrophically forgetting everything before it (cf. §1.2).

[3]I mean here flexibly combining representational primitives to reason and plan multimodally in novel situations well outside of the original training distribution. Formally, statistical learning theory [163] is only concerned with generalization to new samples from the *same* distribution. However, a core aspect of intelligence is autonomous, open-ended generation of one's abstractions to enable adaptation beyond previously experienced training data.

[4]Both autonomously combining motor primitives to learn new compositional skills is currently challenging in robotics, as well as the general principles of sample-efficient, high-dimensional motor control (cf. §1.4).

[5]For example, are explicitly object-centric vision architectures needed (§1.2.1), or are the current architectures mostly sufficient if they had high-dimensional tactile feedback (§1.1.2)? This is all to illustrate that in order to know what's needed for one brain area, it's helpful to have a handle on how other brain areas *interact* with it. Otherwise, when working in isolation, we may be trying to solve much harder problems in one component, that are more easily solved once you integrate these components together.

[6]e.g. visual [173, 73, 25, 115, 180, 119, 118], auditory [72, 44], motor [151, 103], memory [117], and language [141] brain areas.

[7]This quote is famously attributed to the physicist Wolfgang Pauli [130].

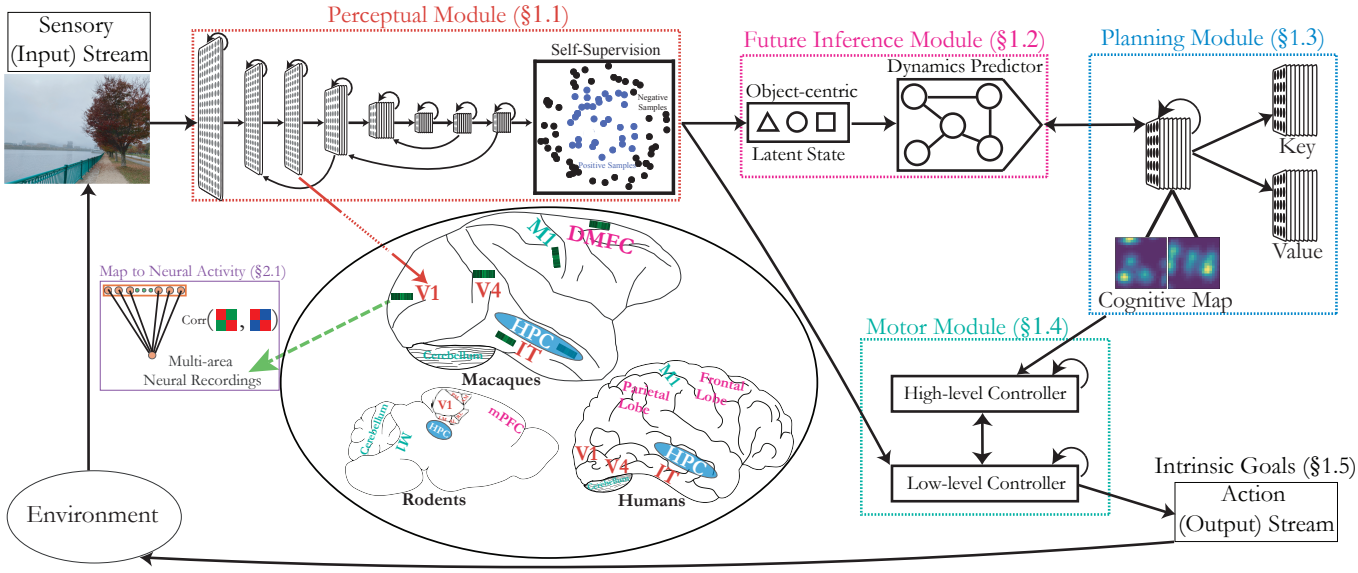

Figure 1: **Integrative, embodied agents to reverse-engineer natural intelligence.** A schematic of an *example* integrative, embodied agent consisting of a recurrent, self-supervised perceptual system (§1.1) that outputs an object-centric latent upon which a future inference module (§1.2) operates to predict the next state of the environment. The planning module (§1.3) hierarchically organizes these representations to plan future actions, which are then passed to effectors that output intrinsically-guided (§1.5) motor commands to perform actions in a biomechanically-realistic animal body (§1.4). Solid black arrows represent *possible* connections between modules. Each representation in these modules is obtained through task-optimization and then mapped (up to the suitable transform, cf. §2.1) to neural activity across multiple brain areas (dotted green arrow). In this example, we show rhesus macaque, rodent, and human brains, with proposed matched representative areas color-coded to each module across species. While I expect the specifics of the modules in each integrative agent to differ for each species it is compared to, the long-term, overarching goal of this approach is that by comparing integrative agents to multi-area neural and behavioral data from multiple species, we are positioned to identify *common* algorithms of natural intelligence conserved across species.

(1)

1. They can **represent/perceive** their physical environment, (§1.1)

2. Which they **act** in, either *reactively* or, (§1.4)

3. By **prediction** of upcoming future states of the environment, (§1.2)

4. Using those predictions to **plan** possible actions to take, (§1.3)

5. According to temporally-varying **intrinsic goals** that are either *hard-wired* or *adaptive* based on the current environmental context (§1.5).

The projects listed in §1 are organized around these five components. *I am, of course, open to considering more components, but I'm focusing on these five to give us a concrete starting point to generate testable hypotheses.* §§2-3 primarily are concerned with applications once we have reasonably developed the embodied agents proposed in §1.

0.1 Methodology

Overall, we are leveraging the nexus of two advances: **(1)** advancements in AI to generate functional hypotheses about the brain, and **(2)** an increasing proliferation of large-scale neural population recordings from various organisms, to strongly constrain our functional hypotheses. This motivates the following methodology we will take when building agents and comparing them to neural and behavioral data, is

applicable across all the lab's projects:

1. **We will assume, as a first-order approximation, that the brains of many animals can be partitioned into functional *modules*.** End-to-end reinforcement learning (RL) has been the latest framework in which agents have been presented. But this framework represents a quite simplistic picture of a brain as a monolithic policy π that gets external rewards from the environment. These rewards are either sparse, or unclear how an actual animal would get these, typically inefficient at controlling even simple multi-jointed bodies [56], which would be deadly for an organism in reality. Now, it could certainly be the case that with enough data, a suitably-chosen survival-based reinforcement objective could recapitulate the brain. But given the data inefficiency and hyperparameter sensitivity of end-to-end RL techniques, we lack a large enough multimodal dataset to fully enable this. It also may never be feasible, as RL is underpowered, being conceptually akin to simple associative trial-and-error learning (though §2.3 explores some improvements in the context of evolutionary searches).

   But practically speaking, current training datasets have many useful priors, so why not leverage self-supervised pretraining to make learning easier rather than starting from scratch each time? Not to mention, for a *fixed* training dataset, compared to self-supervised learned representations, current RL objectives do not give us neurally-aligned visual systems in both rodents [119] and primates [120]. Therefore, except for §1.4 and §1.5 where supervised/RL pretraining may be more suitable for action-related modules, we will likely use self-supervised pretraining for each module[8]. We can think of this pretraining stage as an efficient proxy for learning representations that were settled on over the course of much longer evolutionary timescales for animals. While AI has reasonable techniques for self-supervised representation learning, using this to pretrain modules is likely not enough on its own to replicate animal embodied intelligence. In fact, we can think of the *agent architecture*, which specifies how these modules *interact*, as corresponding to learning over the organism's lifetime (cf. §2.2) – enabling it to generalize to novel, *out-of-distribution* scenarios at "test time" that it can encounter daily, to ensure its survival. Constructing these general-purpose agent architectures is a major open question in AI, and something that the brain sciences can help guide.

2. **We specify the details of these components in the language of task-optimized modeling, from which we generate an understanding of the functional constraints of the target brain area by evaluating models against neural/behavioral data as *ground truth*.** The conceptual insight we get from task-optimized modeling is a *structural* and *functional* understanding of the selective pressures on the brain area under study, over evolution and development to reach the adult state[9]. Thus, the responses of hundreds of neurons (from hundreds to thousands of stimuli conditions) are effectively *distilled* into an *architecture* and optimized *loss function* that faithfully generates it. Therefore, if a lot of models match the data that's bad, because it means the data (and metric) isn't differentiating. But if a small number of them do, then you can learn something, and get a strong conceptual conclusion out of that. Now, we aren't claiming that evolution is always maximizing, or that the brain is always optimal, or that natural selection explains everything in biology by itself. Rather, insofar as some features of neural

[8]For practical purposes, in the event that coming up with self-supervised loss functions for a given module becomes a great impediment (e.g. as it used to be for object categorization, and currently is the case for optic flow), it's fine to use a *supervised*/language-conditioned proxy for the time being, with the eventual goal of likely replacing it with self-supervised loss function(s) later down the line.

[9]There have been recent efforts on "foundation models for neuroscience" [177], achieved by training deep neural networks *directly* on multi-region neural data (rather than optimizing for a task). While we can now collect sufficient neural data without much overfitting when training on it, we avoid this direct neural data training to preserve the understanding of selective pressures that we get through task-optimization, which would otherwise be lost despite yielding a predictive model of neural responses.

response patterns arise through selective pressures under a task constraint, then a task-optimized model of that system will be helpful in explaining those very features (cf. [27] for a more extensive discussion). The past decade of task-optimized modeling has shown that this situation is quite common across species and brain areas. In fact, a core objective of neuroscience has been to uncover the underlying reasons why the structures of the brain are as they are, and why they respond as they do, namely to assign function to structure, which is what we are doing here.

Now, in practice, modern AI changes at a fast pace. The architecture of this year may not be the architecture that's popular in a couple months. Our goal isn't to ship a product or to always beat the most popular machine learning benchmark, but to pursue a persistent, long-term goal of providing a systems-level understanding of an organism's brain. Why? Because the brain is the only example of intelligence we know of and agree on, and over the past century, we've yet to settle on the algorithmic principles that replicate many of the intelligent behaviors it exhibits. As a result, we're not creatively limited to what's efficient for the current hardware of our time (e.g. Transformers [164]), but rather we're in the business of figuring out what tasks and architectures best recapitulate observed neural activity, pulling insights not just from AI, but from neuroscience (mainly for macroscale architecture insights) and cognitive science (mainly for task insights) as well. For example, it's likely inevitable that we will use RNNs in some form since, despite them not being super efficient for GPUs, neurons have functionally-relevant dynamics that our models have to engage with in order to understand their role [118, 115]. Therefore, we're not competing with industry-based AI. Instead, it's a beneficial collaboration, as an algorithm that works robustly for engineering purposes could be a reasonable hypothesis for an algorithm in the brain that we could test[10]. To do so, at a *minimum*, we adapt current AI algorithms to the constraints of the organism under study, rather than only taking pretrained ones off the shelf and comparing them to data. It's this type of "convergent evolution" between science and engineering that, when it happens, can be quite striking and deep.

Now, currently in industry, due to 2-5 year value proposition timelines, it's more practical to scale and curate datasets to get performant models, rather than focus on architectures or loss functions, which might enable more sample-efficient scaling if improved[11]. Especially considering the resource limitations of academic labs, we're certainly more incentivized to spend time inventing more sample-efficient AI algorithms (which I'd argue is the fun, creative part of AI anyhow!). The great advantage of academia is the freedom and luxury to work on the *next* big thing, rather than the current one, and to have the time to incubate and develop it before it's ready for industry adoption. As a result, discovering these inductive biases is what we're best positioned to work on, by thinking both cognitively about the organism's behavioral constraints, and suitably abstracting neural circuit motifs to enable more sample-efficient learning [101, 175]. Ideally, these inductive biases scale with more data, but they mainly have to be sufficient for the compute/data resources *at hand*, available through interaction in most environments (which is the context the brain operates in).[12] From a practical engineering perspective, improving the existing architectures and loss functions in this way may ultimately end up being necessary, as we may end up not having enough training data to scale the current Transformer architectures with

[10]Likewise, if an approach *fails* to work in industry, despite having vast amounts of people and resources working on it, then that is super useful to know. In fact, it's probably the strongest empirical "steelman" one could have *against* a given approach, aside from any yet to be proven no-go theorem.

[11]e.g. just as CNNs work better than MLPs on ImageNet, due to CNNs having more appropriate inductive biases for vision [38, 76], or LSTMs [63] being better at sequence learning tasks over less constrained SimpleRNNs [43], etc.

[12]To illustrate this point further, I'll analogize to existing technology: You could say that abstracting the Bernoulli principle from bird flight enabled us to build jet planes that can transport large amounts of cargo over very long distances. But, compared to a bird, a jet plane is far less flexible in terms of the flight paths it can maneuver or the places it can land on (e.g. planes can't land on tree branches, or fly between tight obstacles). So clearly there is a resource-tradeoff curve, where the bird is optimal in one resource regime, and the plane is optimal for another.

autoregressive loss functions to human (or animal-level) embodied intelligence[13]. And even if we *do* get sufficient interaction data to pretrain statistical representations with existing models, such that most of the evaluations basically become in-distribution at test time, you could still argue that a key essence of intelligence is to be able to perform well in *out-of-distribution* contexts. This makes it necessary to find suitable *agent architectures* that specify how these statistically pretrained components *interact*, to extrapolate far beyond the pretraining data distribution. *After all, it would be infeasible for evolution to pre-program every new situation an animal will come to encounter in its daily life into its genome!* Instead, it needed strong inductive biases to generalize *not* to *all* scenarios, but specifically to those that matter for survival.

3. **For the most part, we leverage high-variation, virtual reality environments for model training and evaluation.** No single robotics simulator is perfect, but as it stands now, current robot hardware is quite inflexible compared to animal bodies and easy to break (allowing us less opportunities to fail). While robot hardware is improving, virtual reality (VR) has a much faster iteration time than the real world, enabling us to have greater control over the environment and to use more biomechanically-realistic animal bodies. Science progresses when there is good control over variables of manipulation, and a fast time to failure! Simulation environments also better line up with evaluating against modern neuroscience experiments which themselves are also using VR increasingly. We *do* eventually want to try controlling real robot bodies, after sufficient iteration in VR. In other words, we test our theories of intelligence with both a "robot test" in collaboration with roboticists, as well as "neural and behavioral tests" in collaboration with neuroscientists and cognitive scientists. Our ultimate engineering endpoint is to provide the software/algorithms for more physically-grounded, common-sense robots, and unlike any single machine learning benchmark, the strongest (and most constraining) test of our theories of intelligence will be if they can empower meaningful interaction with the real world, just as an organism's brain would.

0.2 Philosophy

This endeavor naturally requires a combined interaction between neuroscience, cognitive science, and AI, paving the way towards a more unified *natural science of intelligence*[14]. But what would such a science even look like? While we currently view these as distinct disciplines, they are all different sides of the same coin: (systems) neuroscience is concerned with neural circuit implementations/mechanisms related to a particular function/behavior, cognitive science with those functions/behaviors, and AI with building functional systems[15]. With this perspective in mind, it becomes clear that at a *minimum*, such a science must consist of *functional* and *predictive* theories of natural intelligence, by which I mean: these theories have to be predictive, by recapitulating observed neural activity (neuroscience) and behavioral patterns, while also adequately performing a function/behavior (cognitive science). This will require significant engineering (AI) efforts, in order to scale to the challenging environments animals naturally confront. As a consequence, the "principles" of the natural science of intelligence will mainly be *algorithms*, processes that are abstracted from biology and runnable across different implementations[16].

[13]This possibility stems from the observation that existing generative AI notoriously struggles to reason or plan as well as humans and animals, especially in novel scenarios [34, 70]. Augmenting them further with agent modules (as surveyed here [166]) are unfortunately too application-specific, rather than general-purpose. Brains are also more power-efficient [59] than current AI architectures in many respects (again, likely because of the inductive biases they possess).

[14]I want to distinguish this from the "science of natural intelligence", as the brain and cognitive sciences are pursuing this already, without necessarily having the goal of contributing to, or relying on, insights from building AI that functions in the real world.

[15]Simply put in the language of task-optimized modeling: neuroscience mainly tells us about the *architecture* (through anatomical observations), cognitive science about the *tasks* (consisting of a loss function and data stream), and current AI gives us effective *learning rules* (via gradient-based optimization).

[16]Just as how physical laws are runnable abstractions of physical systems. After all, one could record patterns of voltages from a computer while it runs a program, just as we do in the brain, but that would be completely missing the forest for the trees – the algorithm itself is the understanding we wanted! Put another way, the algorithm sets the *target*

Thus, this science should be applicable not just to machines and human brains, but to other species' nervous systems as well. Finally, I would argue that we are mainly in the "*empirics gathering*" stage of the natural science of intelligence, wherein we are trying to engineer/discover examples of artificial systems that can work in the real world, and can also predict biological brain data. Almost every science begins this way, by collecting many empirical examples of a given phenomenon before being in a position to identify unifying principles (if any) responsible for the common trends that have been observed.

The above considerations lead me to our lab's scientific philosophy:

1. **We are chiefly motivated by a *behavioral* domain of interest.** Specifically, for each species we study, we focus on the hard engineering problem their brain has solved, better than any AI system we have now. And due to Moravec's paradox [109], there are many more of these challenges as we "climb up" the phylogenetic tree, compared to focusing on higher-level cognition, which is easier to reverse-engineer with current AI and has evolved over a much shorter period. While "simpler" or "more complex" organisms aren't objectively defined, in practice, "simple" refers to those species most accessible in terms of neuron count and temporal fidelity using current neural recording technologies. Thus, we work "bottom up"[17], starting with the "simplest" species whose brain solves that hard problem, since it makes the reverse-engineering of that hard problem more tractable compared to a more complex organism (and the insights we find in the simpler setting may have been conserved or independently rediscovered by evolution, serving as an initial hypothesis in another organism). As a result, we do not overfit to any one organism, but through extensive experimental neuroscientist collaborations, we aim to gain an understanding of the algorithmic principles of natural intelligence *across* species. After all, this is the target of explanation for any meaningful natural science of intelligence. In fact, this is a primary reason why we focus more on model building than metric building (though we do some of that[18]) because, for many problems, *no* current model explains the data well under any reasonable metric, as they can't perform the behavior to begin with.

2. **We are interested in theories that are both *functional* in that they do things as an organism would, and are also *predictive* of neural and/or behavioral data.** After all, we are doing *science*, not *only* engineering, and it is a basic requirement of any branch of science that its theories explain empirical data. Predictivity has actually been a high bar in the brain sciences. Many mathematically interpretable hypotheses that were not task-performant, did not end up being predictive of large-scale neural activity[19]. Perhaps this lack of mathematical interpretability is not surprising, considering that a primary observation of modern complex systems research is that in Nature and *in silico* [171], many repeatedly iterated processes (e.g. via evolution or optimization) become much more complex by the end of that iteration[20]. Furthermore, if the past half-century of AI has taught us anything, it's that we know very little about what works

for what to look for in detailed neural response patterns.

[17]This approach is consistent with Gall's Law: "All complex systems that work evolved from simpler systems that worked." A quote due to Rodney Brooks [14, pg. 50] that illustrates the importance of Gall's Law is that if you start by reverse-engineering a plane using a Boeing 747 as your exemplar, you might be too distracted by the other features, like the plastic seats or windows, to realize that flight is due to the wings: "Suppose it is the 1890s. Artificial flight is the glamor subject in science, engineering, and venture capital circles. A bunch of [artificial flight] researchers are miraculously transported by a time machine to the 1990s for a few hours. They spend the whole time in the passenger cabin of a commercial passenger Boeing 747 on a medium duration flight. Returned to the 1890s they feel invigorated, knowing that [artificial flight] is possible on a grand scale. They immediately set to work duplicating what they have seen. They make great progress in designing pitched seats, double pane windows, and know that if only they can figure out those weird 'plastics' they will have the grail within their grasp."

[18]Over the years, a method we've found useful is to first identify the simplest transform between two animals for a given brain area, giving rise to the "inter-animal consistency", and then use that transform for each of the models we compare. See, e.g. [120, 119, 117] for more details. The project in §2.1 is devoted to developing this further.

[19]e.g. Gabors and macaque V1 [25], HMAX and macaque V4/IT [173], grid-cell-only models and rodent MEC [117].

[20]In fact, only a small class of functions, known as *idempotent* functions, stay *identically* the same at every step of the

computationally and why. Many *a priori* theoretically well-motivated ideas have failed to produce intelligent behaviors in the complex environments that organisms face. As such, we will always compare our models against a *range* of alternative hypotheses without being beholden ahead of time to any particular one. The insight comes when many of the hypotheses are non-trivially ruled out by neural data, as none of them were strawmen to begin with. Finally, when possible, we compare to the randomly initialized architecture (or "randomly moving" control), to facilitate understanding and isolate how much individual factors such as the task vs. the architecture alone matters.

3. **Relatedly, we do not *a priori* select for the mathematical interpretability of models (and metrics), but rather choose the most *parsimonious functional* solution that is equally *predictive* at the suitable biological abstraction.** The brain was not specifically evolved to be mathematically beautiful when examined by humans millions of years later, but is the result of many complex and interacting pressures to ensure the organism's survival. *In other words, evolution had to ultimately solve a hard, and constraining, engineering problem.* This is why toy models and tasks, traditionally preferred in computational neuroscience for their mathematical interpretability, often don't suffice for us. They can be too reductionist to exhibit interesting behaviors and too underconstrained, being solvable in many different ways. After all, in order to understand what computational ingredients are needed to *generate* intelligence, our models have to demonstrate these behaviors in the first place! Mathematical interpretability is also an inherently subjective measure[21], whereas *predictivity* of empirical data is an objective measure of scientific progress. This does not mean that in practice we expect to always explain 100% of our data. Rather, we always want to progressively rule out prior theories and find what is most consistent with data with the best engineering solutions we can currently come up with. In other words, we strive to make our theories precise and predictive, so that we can be (slightly) *less wrong* and know by how much they are wrong, rather than being *not even wrong*.

   Task-optimized models, therefore, naturally engage with this complexity by *not a priori* requiring that the *end state* of an evolved/iterated system (like the brain) be mathematically interpretable in all cases. Rather, they provide a mathematically concise loss function and architecture that, once *optimized*, gives rise to the complexity/diversity of neural response profiles we see in a macroscopically (and functionally) significant brain area. Now, these representations may end up being "mere" statistical approximations to a much lengthier implementation-level/mechanistic description of how the parts within a single module/brain area combine into a whole[22]. The main

iteration (e.g. absolute value, $|x|$). Of course, this does not preclude that small portions of the end states of an iterative process *can* remain mathematically interpretable, e.g. as in Ramsey theory [50] where one can prove that if a graph is large enough, then some property P holds in its substructure. So perhaps it's not all that surprising as we see in neural population data that these tend to form a small subset of the population response profiles (e.g. grid cells in MEC, or Gabor edge detectors in V1).

[21] Even if interpretability is subjective, one could still argue that whatever definition of interpretability satisfied someone, that it is much easier to attain this in a task-optimized *in silico* system that they have perfect access to, than the much noisier, limited access biological system (see, e.g. [100, 154, 94], for how this can be achieved in artificial neural networks (ANNs) relevant to neurobiology). For if they cannot attain the former, it is hard to argue they could ever attain the latter. And if such a deeper interpretable theory exists, then we can view what we are doing as the necessary "empirics gathering" stage to strongly *separate* hypotheses about what such a theory could even look like, through the study of successfully brain-predictive ANNs. This is especially the case since it is currently faster to build an ANN is task-performant and predicts a brain area, than to theorize *ab initio* about it and hope it explains the brain.

[22] Although, I will point out that humans have historically been quite bad at providing these handcrafted mechanistic explanations, which one could refer to as the analogous "bitter lesson" [152] of systems neuroscience. In fact, the notion of "mechanism" in traditional neurobiology is a bit ill-defined in a *general* sense, because you end up with a different explanation for what a given brain area is doing for each different input/environment [154], and it is unclear how to combine these mechanisms by hand into a comprehensive account [114]. Instead, it's been more promising to go in the *reverse* direction, by first building the *single* network to natural scene responses and then distilling the individual mechanisms that arise in any specific context (see, e.g. [154, 94]).

point is that such approximations are sufficiently predictive and can be described *succinctly* into the three-tuple of *task* (consisting of a loss function and data stream), *architecture*, and *learning rule*. The core insight is when only a few patterns of tasks and architectures predict the data well, and the rest fall by the wayside, constituting an empirically faithful *distillation* of a complex system. Thus, in some sense (namely, the Kolmogorov one[23]), this would be *more* interpretable than the much longer mechanistic description we may never find.

4. **We are interested in biological features only *as necessary* for improved function, to generate a normative understanding of their role in performing a behavior.** The brain is the most complex biological system in existence, and there are many aspects of it that may not be relevant to intelligence. Therefore, we never add biological realism merely for the sake that the brain has it, but rather we want to know the *minimal* set of details needed for the operationalized behavior under study. *After all, science is, in large part, about finding runnable abstractions ("laws") of a natural phenomenon that are useful for predicting it in new circumstances.* (This is also why we don't train on neural data directly, nor is our goal to do full-brain emulation at the level of molecular biophysics.) In fact, it is a deeper result if we absolutely *fail* without adding extra biological features, than if we add them without understanding their fundamental importance in meaningfully contributing to behavior. A modern example of this point is that having (real-valued) McCulloch-Pitts neurons [99] ended up being a *necessary* abstraction of biological neurons to produce more flexible AI[24], since operating only at the level of behavior (e.g. as operationalized by logic-based symbolic programs in the GOFAI era) ended up having limited success on its own [92]. For the most part, many questions we will ask mainly engage with the level abstraction of *firing rates* (or, slightly less ideally, a related readout of population activity, such as calcium imaging or voxel responses). This isn't merely an aesthetic choice: the last decade of task-optimized modeling has shown empirically, across brain areas and species, that neurons (at the level of their firing rates) are strongly constrained by behavior (as operationalized by optimization on a high-variation, ethologically-relevant task). In words, firing rates are likely a reasonable abstraction for studying natural intelligence in many species and brain areas[25].

[23] `https://en.wikipedia.org/wiki/Kolmogorov_complexity`

[24] We could further ask if we need to incorporate additional biological features beyond rate-based ANNs to replicate natural intelligence, such as dendrites, neurotransmitters, etc. While detailed biophysical modeling is likely necessary for building improved brain-machine interfaces (cf. §3.2), I'm not convinced this is absolutely necessary for intelligence itself (but certainly open to the possibility!). This is due to the "separation of scales", whereby complex phenomena in Nature often emerge collectively over aggregates of finer details. (For example, we never need to resort to quantum mechanics when reasoning about the physics of everyday objects.) So perhaps firing rates aggregated over large neural populations are a sufficient level of description for most of the intelligent behaviors we care to study. This is underscored by the observation that *groups* of artificial units are often needed to faithfully map to one biological neuron, whereas one-to-one transforms routinely fail to do this well (see, e.g. [119, 117]). Therefore, if there is functional relevance to these other biological features, we may be able to "make up for them" by having many more *deeper* [13] or more *stateful*, rate-based recurrent architectures, e.g. LSTMs [63], GRUs [31], differentiable key-value memories [53, 150, 8], and most recently, SSMs [55]. In fact, some of my earliest work showed that a suitably-optimized LSTM architecture is a reasonably good end-to-end differentiable implementation of the readily releasable pool of vesicles [100, §3.6], which is the underlying mechanism for contrast adaptation in the retina [126].

[25] If, on the other hand, we were dealing with questions of timing or energy efficiency, especially at the sensory periphery, it may be necessary to compare to individual spikes. However, questions of energy efficiency are often best addressed *not* by adapting the underlying algorithm, but rather the *hardware* to improve the demands of the software that it runs. Furthermore, it's been a challenge to build spiking models that can perform interesting behaviors, especially compared to rate-based networks. While this situation may change, most of our questions deal with basic function, so firing rates will nonetheless be a sufficient abstraction for the projects considered here. It's an interesting question whether one can come up with an automated "compilation procedure" for going from firing rates to spikes, just as we compile higher-level programming languages to hardware-specific byte code and do all of our reasoning/software development in the general-purpose, high-level language. Perhaps a "population factor" approach [39] might be a good starting point for designing such an automated procedure that translates performant firing rate models into similarly-performant spiking models, which are in turn energy-efficient on specific neuromorphic hardware (effectively becoming a "mortal computation" [61]).

1 Agent Development Projects

In what follows, I now describe projects that cut across species, aiming to address systems-level questions about organisms that are also important for the design of embodied agents. The subsections are organized according to the five basic components of an agent (1), where each project is motivated by identifying suitable architectures/tasks that specify these components, so that they can be naturally combined together to explain neural activity in a variety of species, enabling cross-species comparisons as well. Where relevant, I will highlight the connections between the components, and the species whose brain data we quantitatively compare to for each project in blue.

The next 14 pages detail specific lab projects (omitted). If this type of work interests you, email me to learn more specifics about them, or to collaborate!

References

[1] Brett R Aiello, Kathryn E Stanchak, Alison I Weber, Tanvi Deora, Simon Sponberg, and Bingni W Brunton. Spatial distribution of campaniform sensilla mechanosensors on wings: form, function, and phylogeny. *Current Opinion in Insect Science*, 48:8–17, 2021.

[2] Diego Aldarondo, Josh Merel, Jesse D Marshall, Leonard Hasenclever, Ugne Klibaite, Amanda Gellis, Yuval Tassa, Greg Wayne, Matthew Botvinick, and Bence P Ölveczky. A virtual rodent predicts the structure of neural activity across behaviors. *Nature*, pages 1–3, 2024.

[3] Kristin K Anstrom, Klaus A Miczek, and EA Budygin. Increased phasic dopamine signaling in the mesolimbic pathway during social defeat in rats. *Neuroscience*, 161(1):3–12, 2009.

[4] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[5] Christa A Baker, Claire McKellar, Rich Pang, Aljoscha Nern, Sven Dorkenwald, Diego A Pacheco, Nils Eckstein, Jan Funke, Barry J Dickson, and Mala Murthy. Neural network organization for courtship-song feature detection in drosophila. *Current Biology*, 32(15):3317–3333, 2022.

[6] Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning.

[7] Rahul Bale, Max Hao, Amneet Pal Singh Bhalla, and Neelesh A Patankar. Energy efficiency and allometry of movement of swimming and flying animals. *Proceedings of the National Academy of Sciences*, 111(21):7517–7521, 2014.

[8] Andrea Banino, Adria Puigdomenech Badia, Raphael Köster, Martin J Chadwick, Vinicius Zambaldi, Demis Hassabis, Caswell Barry, Matthew Botvinick, Dharshan Kumaran, and Charles Blundell. Memo: A deep network for flexible combination of episodic memories. *arXiv preprint arXiv:2001.10913*, 2020.

[9] Andrea Banino, Jan Balaguer, and Charles Blundell. Pondernet: Learning to ponder. *arXiv preprint arXiv:2107.05407*, 2021.

[10] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

[11] Daniel Bear, Elias Wang, Damian Mrowca, Felix Jedidja Binder, Hsiao-Yu Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin A Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[12] Daniel M Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel LK Yamins. Unifying (machine) vision via counterfactual world modeling. *arXiv preprint arXiv:2306.01828*, 2023.

[13] David Beniaguev, Idan Segev, and Michael London. Single cortical neurons as deep artificial neural networks. *Neuron*, 109(17):2727–2739, 2021.

[14] Max Bennett. *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains*. Mariner Books, 2023. ISBN 9780063286344.

[15] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

[16] Nikhil Bhattasali, Anthony M Zador, and Tatiana Engel. Neural circuit architectural priors for embodied control. *Advances in neural information processing systems*, 35:12744–12759, 2022.

[17] Johann H Bollmann. The zebrafish visual system: from circuits to behavior. *Annual review of vision science*, 5: 269–293, 2019.

[18] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[19] Nick Bostrom. Ethical issues in advanced artificial intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al.*, pages 12–17, 2003. URL `https://nickbostrom.com/ethics/ai`. A slightly revised version available at `https://nickbostrom.com/ethics/ai`.

[20] Bella E Brezovec, Andrew B Berger, Yukun A Hao, Feng Chen, Shaul Druckmann, and Thomas R Clandinin. Mapping the neural dynamics of locomotion across the drosophila brain. *Current Biology*, 34(4):710–726, 2024.

[21] Rodney Brooks. A robust layered control system for a mobile robot. *IEEE journal on robotics and automation*, 2 (1):14–23, 1986.

[22] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[23] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[24] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

[25] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[26] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosuite–a contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*, 2022.

[27] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85:101200, 2024.

[28] Megan R Carey. The cerebellum. *Current Biology*, 34(1):R7–R11, 2024.

[29] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[30] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607.

[31] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL `https://www.aclweb.org/anthology/W14-4012`.

[32] Jeff Clune. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*, 2019.

[33] Philip Coen, Marjorie Xie, Jan Clemens, and Mala Murthy. Sensorimotor transformations underlying variability in song intensity during drosophila courtship. *Neuron*, 89(3):629–644, 2016.

[34] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022.

[35] Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792): 671–675, 2020.

[36] Saskia EJ de Vries, Jerome A Lecoq, Michael A Buice, Peter A Groblewski, Gabriel K Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. 23(1):138–151.

[37] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[39] Brian DePasquale, David Sussillo, LF Abbott, and Mark M Churchland. The centrality of population-level factors to network computation is demonstrated by a versatile approach for training spiking networks. *Neuron*, 111(5): 631–649, 2023.

[40] Aniket Didolkar, Anirudh Goyal, and Yoshua Bengio. Cycle consistency driven object discovery. *arXiv preprint arXiv:2306.02204*, 2023.

[41] Ruxu Du, Zheng Li, Y Kamal, and V Pablo. Robot fish. *Springer Tracts in Mechanical Engineering*, 2015.

[42] Howard Eichenbaum, Paul Dudchenko, Emma Wood, Matthew Shapiro, and Heikki Tanila. The hippocampus, memory, and place cells: is it spatial memory or a memory space? 23(2):209–226.

[43] Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[44] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In *Advances in Neural Information Processing Systems*, pages 10078–10089.

[45] Dawn Finzi, Eshed Margalit, Kendrick Kay, Daniel LK Yamins, and Kalanit Grill-Spector. A single computational objective drives specialization of streams in visual cortex. *bioRxiv*, pages 2023–12, 2023.

[46] Jason Fischer, John G Mikhael, Joshua B Tenenbaum, and Nancy Kanwisher. Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, 113(34):E5072–E5081, 2016.

[47] Valerio Francioni, Vincent D Tang, Norma J Brown, Enrique HS Toloza, and Mark Harnett. Vectorized instructive signals in cortical dendrites during a brain-computer interface task. *bioRxiv*, 2023.

[48] Samuel J. Gershman, John A. Assad, Sandeep Robert Datta, Scott W. Linderman, Bernardo L. Sabatini, Naoshige Uchida, and Linda Wilbrecht. Explaining dopamine through prediction errors and beyond. *Nature Neuroscience*, 2024. ISSN 1546-1726. doi: 10.1038/s41593-024-01705-4. URL `https://doi.org/10.1038/s41593-024-01705-4`.

[49] Charles D. Gilbert and Li Wu. Top-down influences on visual processing. *Nat. Rev. Neurosci.*, 14(5):350–363, 2013.

[50] Ronald L Graham, Bruce L Rothschild, and Joel H Spencer. *Ramsey theory*, volume 20. John Wiley & Sons, 1991.

[51] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[52] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[53] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

[54] Stephen Grossberg. Competitive Learning: From Interactive Activation to Adaptive Resonance. 11:23–63.

[55] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[56] Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Thomas Paine, Sergio Gómez, Konrad Zolna, Rishabh Agarwal, Josh S Merel, Daniel J Mankowitz, Cosmin Paduraru, Gabriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matthew Hoffman, Nicolas Heess, and Nando de Freitas. RL unplugged: A suite of benchmarks for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 7248–7259, 2020.

[57] Martin Haesemeyer, Alexander F Schier, and Florian Engert. Convergent temperature representations in artificial and biological neural networks. *Neuron*, 103(6):1123–1134, 2019.

[58] Julie A Harris, Stefan Mihalas, Karla E Hirokawa, Jennifer D Whitesell, Hannah Choi, Amy Bernard, Phillip Bohn, Shiella Caldejon, Linzy Casal, Andrew Cho, et al. Hierarchical organization of cortical and thalamic connectivity. 575(7781):195–202.

[59] Gal Haspel, Edward S Boyden, Jeffrey Brown, George Church, Netta Cohen, Christopher Fang-Yen, Steven Flavell, Miriam B Goodman, Anne C Hart, Oliver Hobert, et al. To reverse engineer an entire nervous system. *arXiv preprint arXiv:2308.06578*, 2023.

[60] Marion F Haug, Oliver Biehlmaier, Kaspar P Mueller, and Stephan CF Neuhauss. Visual acuity in larval zebrafish: behavior and histology. *Frontiers in Zoology*, 7:1–7, 2010.

[61] Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

[62] Noriaki Hirose, Dhruv Shah, Kyle Stachowicz, Ajay Sridhar, and Sergey Levine. Selfi: Autonomous self-improvement with reinforcement learning for social navigation. *arXiv preprint arXiv:2403.00991*, 2024.

[63] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[64] Peter C Humphreys, Kayvon Daie, Karel Svoboda, Matthew Botvinick, and Timothy P Lillicrap. Bci learning phenomena can be explained by gradient-based optimization. *bioRxiv*, pages 2022–12, 2022.

[65] Kristopher T Jensen, Guillaume Hennequin, and Marcelo G Mattar. A recurrent network model of planning explains hippocampal replay and human behavior. *Nature Neuroscience*, pages 1–9, 2024.

[66] Huijeong Jeong, Annie Taylor, Joseph R Floeder, Martin Lohmann, Stefan Mihalas, Brenda Wu, Mingkang Zhou, Dennis A Burke, and Vijay Mohan K Namboodiri. Mesolimbic dopamine release conveys causal associations. *Science*, 378(6626):eabq6740, 2022.

[67] Yusheng Jiao, Feng Ling, Sina Heydari, Nicolas Heess, Josh Merel, and Eva Kanso. Learning to swim in potential flow. *Physical Review Fluids*, 6(5):050505, 2021.

[68] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*, 2018.

[69] Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*, 2023.

[70] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.

[71] Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6): 974–983, 2019.

[72] Alexander JE Kell*, Daniel LK Yamins*, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[73] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[74] Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *International conference on machine learning*, pages 5306–5315. PMLR, 2020.

[75] Leo Kozachkov, Ksenia V Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(34):e2219150120, 2023.

[76] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[77] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[78] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021.

[79] Daniel Kunin*, Aran Nayebi*, Javier Sagastuy-Brena*, Surya Ganguli, Jonathan Bloom, and Daniel Yamins. Two routes to scalable credit assignment without weight symmetry. In *International Conference on Machine Learning*, pages 5511–5521. PMLR, 2020.

[80] International Brain Laboratory. Data release - Brainwide map - Q4 2022. 11 2022. doi: 10.6084/m9.figshare.21400815.v5. URL `https://figshare.com/articles/preprint/Data_release_-_Brainwide_map_-_Q4_2022/21400815`.

[81] John E. Laird, Mazin Assanie, James Kirk, Peter Lindes, Aaron Mininger, and Shiwali Mohan. Rosie itl agent, 2024. URL `https://soar.eecs.umich.edu/rosie/`. Soar Agent that learns through situated interactive instruction in a robotic environment.

[82] Andrew Lampinen, Stephanie Chan, Andrea Banino, and Felix Hill. Towards mental time travel: a hierarchical memory for reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34:28182–28195, 2021.

[83] Janne K Lappalainen, Fabian D Tschopp, Sridhama Prakhya, Mason McGill, Aljoscha Nern, Kazunori Shinomiya, Shin-ya Takemura, Eyal Gruntman, Jakob H Macke, and Srinivas C Turaga. Connectome-constrained deep mechanistic networks predict neural responses across the fly visual system at single-neuron resolution. *bioRxiv*, pages 2023–03, 2023.

[84] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

[85] Susan J Lederman and Roberta L Klatzky. Haptic perception: A tutorial. *Attention, Perception, & Psychophysics*, 71(7):1439–1459, 2009.

[86] H. Lee, E. Margalit, K. M. Jozwik, M. A. Cohen, N. Kanwisher, D. L. Yamins, and J. J. DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 07 2020. URL `https://www.biorxiv.org/content/early/2020/07/01/2020.07.01.183384`.

[87] Jacques C Leedekerken, Maurice F Fallon, and John J Leonard. Mapping complex marine environments with autonomous surface craft. pages 525–539. Springer, Berlin, Heidelberg, 2014.

[88] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.

[89] Chenhao Li, Elijah Stanger-Jones, Steve Heim, and Sangbae Kim. Fld: Fourier latent dynamics for structured motion representation and learning. *arXiv preprint arXiv:2402.13820*, 2024.

[90] Liang Li, Chen Wang, and Guangming Xie. A general cpg network and its implementation on the microcontroller. *Neurocomputing*, 167:299–305, 2015.

[91] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1810.01566*, 2019.

[92] J. Lighthill. Artificial intelligence: A general survey. Technical Report AC 11/2/10, Science Research Council, London, UK, 1973. URL `https://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm`.

[93] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. pages 1–12.

[94] Niru Maheswaranathan, Lane T McIntosh, Hidenori Tanaka, Satchel Grant, David B Kastner, Joshua B Melander, Aran Nayebi, Luke E Brezovec, Julia H Wang, Surya Ganguli, et al. Interpreting the retinal neural code for natural scenes: From computations to neurons. *Neuron*, 111(17):2742–2755, 2023.

[95] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *arXiv preprint arXiv:2303.18240*, 2023.

[96] Jesse D Marshall, Diego E Aldarondo, Timothy W Dunn, William L Wang, Gordon J Berman, and Bence P Ölveczky. Continuous whole-body 3d kinematic recordings across the rodent behavioral repertoire. *Neuron*, 109 (3):420–437, 2021.

[97] Fernando Martínez-García and Enrique Lanuza. Evolution of vertebrate survival circuits. *Current Opinion in Behavioral Sciences*, 24:113–123, 2018.

[98] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[99] WS McCulloh and W Pitts. A logical calculus of the ideas immanent in neural nets. *Bull Math. Biophys*, 5:133–137, 1943.

[100] Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. Deep learning models of the retinal response to natural scenes. *Advances in neural information processing systems*, 29, 2016.

[101] Josh Merel, Matthew Botvinick, and Greg Wayne. Hierarchical motor control in mammals and machines. *Nature communications*, 10(1):1–12, 2019.

[102] Josh Merel*, Diego Aldarondo*, Jesse Marshall*, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. *International Conference on Learning Representations*, 2020.

[103] Jonathan A Michaels, Stefan Schaffelhofer, Andres Agudelo-Toro, and Hansjörg Scherberger. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. 117(50):32124–32135.

[104] Tomáš Mikolov et al. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80(26), 2012.

[105] Marvin Minsky. *Society of mind*. Simon and Schuster, 1988.

[106] Marvin Minsky. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster, 2007.

[107] Marvin Minsky and Seymour A Papert. *Perceptrons, reissue of the 1988 expanded edition with a new foreword by Léon Bottou: an introduction to computational geometry*. MIT press, 2017.

[108] Thomas L Mohren, Thomas L Daniel, Steven L Brunton, and Bingni W Brunton. Neural-inspired sensors enable sparse, efficient classification of spatiotemporal data. *Proceedings of the National Academy of Sciences*, 115(42): 10564–10569, 2018.

[109] Hans Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988.

[110] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

[111] Yu Mu, Davis V Bennett, Mikail Rubinov, Sujatha Narayan, Chao-Tsung Yang, Masashi Tanimoto, Brett D Mensh, Loren L Looger, and Misha B Ahrens. Glia accumulate evidence that actions are futile and suppress unsuccessful behavior. *Cell*, 178(1):27–43, 2019.

[112] Aishwarya Nair, Alejandro Alvaro, and Siddhartha Verma. Navigation of interacting swimmers using multi-agent reinforcement learning. *Bulletin of the American Physical Society*, 2023.

[113] Eva A Naumann, James E Fitzgerald, Timothy W Dunn, Jason Rihel, Haim Sompolinsky, and Florian Engert. From whole-brain data to functional circuit models: the zebrafish optomotor response. *Cell*, 167(4):947–960, 2016.

[114] Aran Nayebi. *A Goal-Driven Approach to Systems Neuroscience*. Stanford University, 2022. URL `https://purl.stanford.edu/qk457cr2641`.

[115] Aran Nayebi*, Daniel Bear*, Jonas Kubilius*, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel L Yamins. Task-driven convolutional recurrent models of the visual system. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[116] Aran Nayebi, Sanjana Srivastava, Surya Ganguli, and Daniel L Yamins. Identifying learning rules from neural network observables. *Advances in Neural Information Processing Systems*, 33:2639–2650, 2020.

[117] Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin Mallory, Gabriel Mel, Ben Sorscher, Alex Williams, Surya Ganguli, Lisa M Giocomo, and Daniel LK Yamins. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[118] Aran Nayebi, Javier Sagastuy-Brena, Daniel M Bear, Kohitij Kar, Jonas Kubilius, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel LK Yamins. Recurrent connections in the primate ventral visual stream mediate a tradeoff between task performance and network size during core object recognition. *Neural Computation*, 34: 1652–1675, 2022.

[119] Aran Nayebi*, Nathan CL Kong*, Chengxu Zhuang, Justin L Gardner, Anthony M Norcia, and Daniel LK Yamins. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, 19, 2023.

[120] Aran Nayebi, Rishi Rajalingham, Mehrdad Jazayeri, and Guangyu Robert Yang. Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes. *Advances in Neural Information Processing Systems*, 36, 2023.

[121] Sujaya Neupane, Ila R Fiete, and Mehrdad Jazayeri. Vector production via mental navigation in the entorhinal cortex. *bioRxiv*, pages 2022–12, 2022.

[122] Sujaya Neupane, Ila Fiete, and Mehrdad Jazayeri. Mental navigation in the primate entorhinal cortex. *Nature*, pages 1–8, 2024.

[123] Nils J. Nilsson. Shakey the robot. Technical Note 323, SRI International, Artificial Intelligence Center, Computer Science and Technology Division, April 1984. URL `https://ai.stanford.edu/users/nilsson/OnlinePubs-Nils/shakey-the-robot.pdf`.

[124] Souzana Obretenova, Mark A Halko, Ela B Plow, Alvaro Pascual-Leone, and Lotfi B Merabet. Neuroplasticity associated with tactile language communication in a deaf-blind subject. *Frontiers in human neuroscience*, 3:953, 2010.

[125] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding.

[126] Yusuf Ozuysal, David B Kastner, and Stephen A Baccus. Adaptive feature detection from differential processing in parallel retinal pathways. *PLoS computational biology*, 14(11):e1006560, 2018.

[127] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[128] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[129] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

[130] Rudolf Peierls. Where pauli made his 'wrong' remark. *Physics Today*, 45(12):112, 1992. doi: 10.1063/1.2809934.

[131] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022.

[132] Ulyana Piterbarg, Lerrel Pinto, and Rob Fergus. diff history for long-context language agents. *arXiv preprint arXiv:2312.07540*, 2023.

[133] RT Pramod, Michael A Cohen, Joshua B Tenenbaum, and Nancy Kanwisher. Invariant representation of physical stability in the human brain. *eLife*, 11:e71736, 2022.

[134] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pages 2847–2854. PMLR, 2017.

[135] Rishi Rajalingham, Aída Piccato, and Mehrdad Jazayeri. Recurrent neural networks with explicit representation of dynamic latent variables can mimic behavioral patterns in a physical inference task. *Nature communications*, 13 (1):1–15, 2022.

[136] Rishi Rajalingham, Hansem Sohn, and Mehrdad Jazayeri. Dynamic tracking of objects in the macaque dorsomedial frontal cortex. *bioRxiv*, 2022.

[137] Yann Roussel, Stephanie F Gaudreau, Emily R Kacer, Mohini Sengupta, and Tuan V Bui. Modeling spinal locomotor circuits for movements in developing zebrafish. *Elife*, 10:e67453, 2021.

[138] Morteza Sarafyazd and Mehrdad Jazayeri. Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441):eaav8911, 2019.

[139] Daniel L Schacter, Donna Rose Addis, and Randy L Buckner. Episodic simulation of future events: Concepts, data, and applications. *Annals of the New York Academy of Sciences*, 1124(1):39–60, 2008.

[140] M Schrimpf, P Mc Grath, and J J DiCarlo. Topographic anns predict the behavioral effects of causal perturbations in primate visual ventral stream it. In *Champalimaud Research Symposium (CRS21)*, 2021.

[141] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[142] Sugandha Sharma, Aidan Curtis, Marta Kryven, Josh Tenenbaum, and Ila Fiete. Map induction: Compositional spatial submap learning for efficient exploration in novel environments. *arXiv preprint arXiv:2110.12301*, 2021.

[143] Sugandha Sharma, Sarthak Chandra, and Ila Fiete. Content addressable memory without catastrophic forgetting by heteroassociation with a fixed scaffold. In *International Conference on Machine Learning*, pages 19658–19682. PMLR, 2022.

[144] Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pages 3135–3144. PMLR, 2017.

[145] Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. pages 1–7.

[146] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550 (7676):354–359, 2017.

[147] Annabelle C Singer, Margaret F Carr, Mattias P Karlsson, and Loren M Frank. Hippocampal swr activity predicts correct decisions during the initial learning of an alternation task. *Neuron*, 77(6):1163–1173, 2013.

[148] Clara Kwon Starkweather and Naoshige Uchida. Dopamine signals as temporal difference errors: recent advances. *Current Opinion in Neurobiology*, 67:95–105, 2021.

[149] Nicholas A Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539), 2021.

[150] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.

[151] David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. 18(7):1025–1033.

[152] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

[153] Richard S Sutton. The quest for a common model of the intelligent decision maker. *arXiv preprint arXiv:2202.13252*, 2022.

[154] Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, and Surya Ganguli. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Advances in neural information processing systems*, 32, 2019.

[155] Seth R Taylor, Gabriel Santpere, Alexis Weinreb, Alec Barrett, Molly B Reilly, Chuan Xu, Erdem Varol, Panos Oikonomou, Lori Glenwinkel, Rebecca McWhirter, et al. Molecular topography of an entire nervous system. *Cell*, 184(16):4329–4347, 2021.

[156] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[157] Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Rosa Cao, and Daniel LK Yamins. Inter-animal transforms as a guide to model-brain comparison. In *ICLR 2024 Workshop on Representational Alignment*.

[158] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[159] Manuel Traub, Sebastian Otte, Tobias Menge, Matthias Karlbauer, Jannik Thuemmel, and Martin V Butz. Learning what and where: Disentangling location and identity tracking without supervision. In *The Eleventh International Conference on Learning Representations*, 2022.

[160] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoč, Yue-Him Wong, Ajay Joneja, and Sai-Kit Yeung. Marine video kit: a new marine video dataset for content-based analysis and retrieval. In *International Conference on Multimedia Modeling*, pages 539–550. Springer, 2023.

[161] Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.

[162] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020.

[163] Vladimir Naumovich Vapnik, Vlamimir Vapnik, et al. Statistical learning theory. 1998.

[164] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[165] Roman Vaxenburg, Igor Siwanowicz, Josh Merel, Alice A Robie, Carmen Morrow, Guido Novati, Zinovia Stefanidi, Gwyneth M Card, Michael B Reiser, Matthew M Botvinick, et al. Whole-body simulation of realistic fruit fly locomotion with deep reinforcement learning. *bioRxiv*, pages 2024–03, 2024.

[166] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18 (6):186345, 2024.

[167] Peter Y Wang, Yi Sun, Richard Axel, LF Abbott, and Guangyu Robert Yang. Evolving the olfactory system with machine learning. *Neuron*, 109(23):3879–3892, 2021.

[168] Zeguan Wang, Jie Zhang, Panagiotis Symvoulidis, Wei Guo, Lige Zhang, Matthew A Wilson, and Edward S Boyden. Imaging the voltage of neurons distributed across entire brains of larval zebrafish. *bioRxiv*, pages 2023–12, 2023.

[169] Sibo Wang-Chen, Victor Alfred Stimpfling, Pembe Gizem Özdil, Louise Genoud, Femke Hurtak, and Pavan Ramdya. Neuromechfly 2.0, a framework for simulating embodied sensorimotor control in adult drosophila. *bioRxiv*, pages 2023–09, 2023.

[170] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 1961.

[171] Stephen Wolfram. Statistical mechanics of cellular automata. *Reviews of modern physics*, 55(3):601, 1983.

[172] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[173] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. 111(23):8619–8624.

[174] Jinyang Yuan, Tonglin Chen, Bin Li, and Xiangyang Xue. Compositional scene representation learning via reconstruction: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[175] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

[176] Hongming Zhang and Tianyang Yu. Alphazero. *Deep Reinforcement Learning: Fundamentals, Research and Applications*, pages 391–415, 2020.

[177] Yizi Zhang, Yanchen Wang, Donato Jimenez-Beneto, Zixuan Wang, Mehdi Azabou, Blake Richards, Olivier Winter, The International Brain Laboratory, Eva Dyer, Liam Paninski, et al. Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution. *arXiv preprint arXiv:2407.14668*, 2024.

[178] Chengxu Zhuang, Jonas Kubilius, Mitra JZ Hartmann, and Daniel Yamins. Toward goal-driven neural network models for the rodent whisker-trigeminal system. 2017:2556–2566.

[179] Chengxu Zhuang, Tianwei She, Alex Andonian, Max Sobol Mark, and Daniel Yamins. Unsupervised learning from video with deep neural embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9563–9572, 2020.

[180] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), 2021.

[181] Weijian Zong, Horst A Obenhaus, Emilie R Skytøen, Hanna Eneqvist, Nienke L de Jong, Ruben Vale, Marina R Jorge, May-Britt Moser, and Edvard I Moser. Large-scale two-photon calcium imaging in freely moving mice. *Cell*, 185(7):1240–1256, 2022.

[182] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*, 2017.