Review

Cellular Scaling Laws in the Mammalian Brain

Fei Chen and Evan Z. Macosko *

Posted Date: 10 February 2026

doi: 10.20944/preprints202602.0767.v1

Keywords: cellular neuroscience; single-cell genomics; spatial genomics; artificial intelligence

Review

Cellular Scaling Laws in the Mammalian Brain

Fei Chen [1,2,*] and Evan Z. Macosko [1,3,4,*]

[1] Broad Institute of Harvard and MIT, Cambridge, MA, USA
[2] Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA
[3] Department of Neurobiology, Harvard Medical School, Boston, MA, USA
[4] Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA
* Correspondence: chenf@broadinstitute.org (F.C.); emacosko@broadinstitute.org (E.Z.M.)

Abstract

Recent whole-brain cell atlases have uncovered a consistent cytoarchitectural feature: the greatest diversity of discrete neuronal cell types resides not in the regions with the most neurons (e.g. the cortex and cerebellum) but rather in deep subcortical structures like the hypothalamus and brain stem. We propose that this discrepancy reflects a fundamental algorithmic division in the vertebrate brain between a "learning subsystem" and a "steering subsystem." The learning subsystem (cortex, striatum, cerebellum) scales via the replication of repetitive modules to maximize computational capacity, analogous to scaling up parameters in machine learning models. By contrast, the steering subsystem (hypothalamus, pallidum, brainstem) scales via diversification of bespoke cell types to encode innate drives and reflexes, functioning as a high-dimensional biological "reward function." This framework explains the divergence in how evolution has influenced these subsystems, and offers a unified lens for understanding brain architecture, the etiology of brain disease, and may also inform model design for artificial intelligence.

Keywords: cellular neuroscience; single-cell genomics; spatial genomics; artificial intelligence

1. High Cellular Complexity in Deep Brain Structures

The enormous functional complexity of the mammalian brain is reflected in the massive diversification of its constituent neurons. Understanding how many distinct types of neurons exist, how they are organized, and most importantly, why such diversity evolved in the first place has been a central focus of cellular neuroscience since the foundational work of Santiago Ramón y Cajal. His meticulous anatomical studies revealed that the mammalian brain is indeed composed of a vast morphological diversity of neurons [1]; over the next century, neuroscientists endeavored to classify these cells, primarily relying on morphology, selective protein expression markers, connectivity, and electrophysiology.

A major region of focus during this period was the retina, a highly physically accessible neural tissue with a clear functional role. There, systematic classification revealed dozens of distinct neuronal cell types organized into precise, stereotyped functional circuits [2,3]. Individual cell types were often found to encode unique, specific neuronal computations, helping to understand the functional role of the retina as a whole [4–6]. The success of these cell typing efforts in the retina fueled expectation that other brain regions–particularly the cerebral cortex, as the seat of higher cognition–would possess similar or greater levels of discrete cellular complexity. Indeed, based upon his work in the retina, Richard Masland estimated that the cortex might contain as many as 1,000 different cell types [7]. The discovery of many highly discrete interneuron populations in the cortex reinforced the view that the neocortex may be composed of extremely high numbers of cell types [8,9], spawning formal community efforts to develop systematic classification schemes for these heterogeneous cells [10].

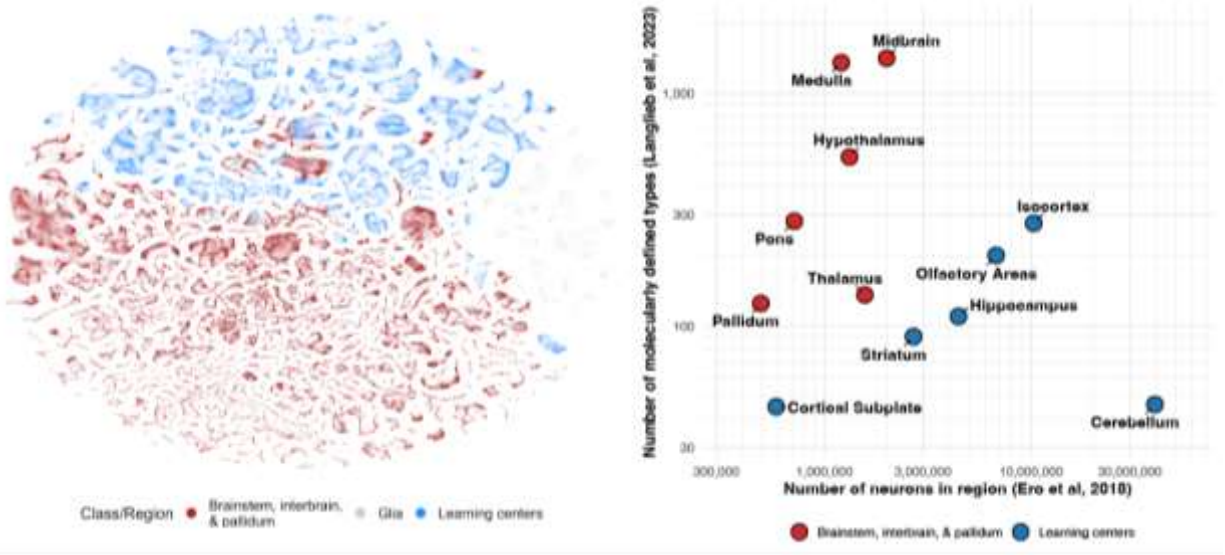
 doi:10.20944/preprints202602.0767.v1

Outside of the cortex, however, detailed neuroanatomical studies hinted at extraordinary complexity within deep subcortical regions. Meticulous morphological and connectivity mapping of hypothalamic and limbic areas, exemplified by the studies of the bed nucleus of the stria terminalis by Larry Swanson, Hongwei Dong, and colleagues [11–13], implied extreme levels of neuronal specialization in these areas. Yet, defining boundaries between cell types based exclusively upon classical features remained challenging [14]; recognizing these limitations, enthusiasm built for utilizing high-throughput genomic methods as a systematic approach for categorizing cells across the brain [15].

Technological advancements in the throughput of single-cell and spatial genomics measurements made the realization of this vision possible. The development of single-cell RNA-sequencing (scRNA-seq) provided an unbiased method to classify cells by their entire transcriptome [16]. In particular, droplet-based methods [17,18], especially applied to nuclei that could be easily extracted from complex neural tissues without enzymatic dissociation [19], dramatically increased scale, enabling the profiling of millions of cells at once. Concurrently, advances in spatial transcriptomics, utilizing both imaging-based strategies [20,21] and sequencing-based indexing [22,23], allowed researchers to map these molecularly defined cell types back to their precise neuroanatomical locations.

These technological breakthroughs recently culminated in the generation of the first comprehensive molecular atlases of the entire mouse brain [24–27]. Systematic sampling of all brain regions identified thousands of molecularly distinct populations. The results converged on a surprising organizational principle, hinted at by earlier suggestions of hypothalamic complexity [21,29,30]: the greatest diversity of neuronal cell types resides in deep subcortical regions within the brain stem, interbrain (hypothalamus and thalamus), and pallidum (**Figure 1**). In addition, the nature of molecular diversification in these different areas is qualitatively different. In these atlases, the subregions of the major learning centers–namely the cerebral and cerebellar cortices, striatum, and hippocampus–constitute duplications of largely the same cell types, with only modest local variations. By contrast, most subnuclei in the brain stem and hypothalamus are composed of highly discrete, bespoke cell types. This massive expansion of specialized cell types in regions associated with innate drives and fixed behaviors raises the fundamental biological question we address here: why have these areas diversified so extensively relative to the brain's learning centers?

We propose that this distinct cellular architecture is best explained by a deep algorithmic division in vertebrate brain function between two tightly coupled systems [31]: the "learning subsystem" and the "steering subsystem" (**Figure 2**). To accomplish their respective functions, each has evolved different, distinct cell type scaling laws.



 doi:10.20944/preprints202602.0767.v1

Figure 1. Left, a t-distributed stochastic neighbor embedding (tSNE) of individual cells from a recent whole-brain atlas (Langlieb, Sachdev *et al*, 2023 [25]). Cells are colored by their region of origin. "Learning centers" include isocortex, hippocampus, striatum, olfactory cortex, the cortical subplate, and the cerebellum. The tSNE algorithm places cells in local proximity to others with shared gene expression; note that red-labeled areas have many more small, distinct clusters compared with blue-labeled areas that have a smaller number of clusters with more cells in each. **Right**, a comparison of the number of neurons quantified in each indicated mouse region by the Blue Brain Cell Atlas [28] (x-axis) and the number of molecular defined cell types in that region.

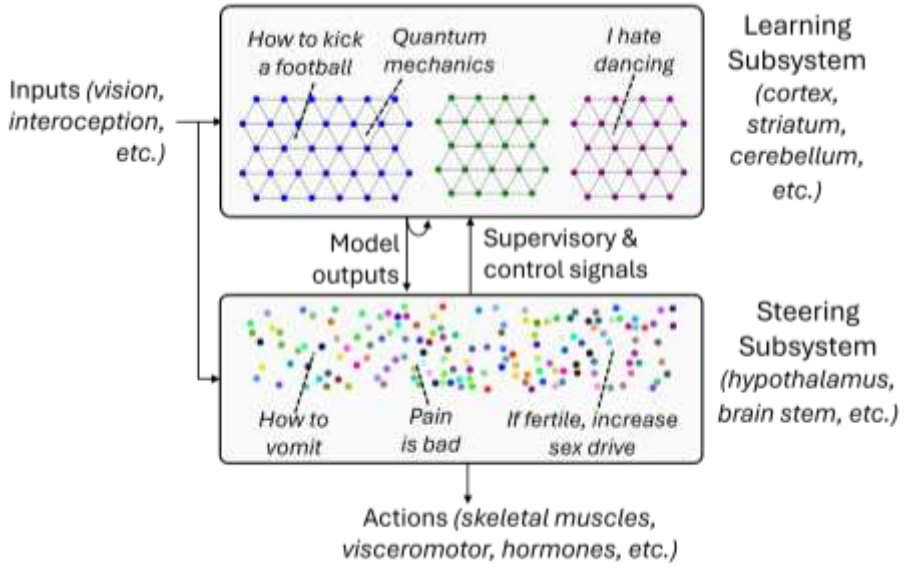


Figure 2. Simplified sketch of our proposed algorithmic architecture for the vertebrate brain. The learning subsystem (top) stores, queries, and updates a small number of scaled-up within-lifetime learning algorithms, which involve many copies of relatively fewer cell types. The steering subsystem (bottom) has a large number of idiosyncratic cell groups, which support innate behaviors, reactions, and drives.

2. The Learning Subsystem: Scaling via Replication

The "learning subsystem" encompasses the cortex, cerebellum, striatum, and hippocampus. These structures are evolutionarily and functionally distinct, yet they share a common architectural logic: they are built from repetitive copies of simpler units, such as the cortical minicolumn, the striatal spiny neuron, or the cerebellar microcomplex.

The primary function of this subsystem is to execute within-lifetime learning algorithms, allowing animals to develop skilled, flexible behaviors adapted to novel environments. These regions are responsible for storing, querying, and updating learned models of the world. Because these algorithms are designed to be general-purpose, the cellular hardware that runs them does not need to be specialized for particular semantic content. Instead, the "content" of the cortex, whether one is looking at a face or listening to music, is mostly determined by the inputs and the training data, not by the constituent cell types themselves. This explains why different parts of cortex can perform radically different functions (e.g. visual processing, math, or language) despite each subregion possessing highly similar cytoarchitecture.

Consequently, the learning subsystem follows **sublinear power-law scaling**: neuron count expands by orders of magnitude to increase storage and information-processing capacity, while the repertoire of neuron types remains relatively constant (**Figure 3**). This is analogous to the "number of model weights" in artificial neural networks [32]. In this framework, the modest cytoarchitectural variations observed across different cortical territories, striatal compartments, or cerebellar lobules–mostly in cell proportions, but with some molecular specializations–reflect evolutionary tweaks to hyperparameters and architectures in learning algorithms to optimize storage of particular kinds of

 doi:10.20944/preprints202602.0767.v1

learned inputs. In machine learning (ML), increasing the capability of a model often involves only "scaling up" the architecture–adding more layers or more neurons–without changing the fundamental algorithm or underlying code that defines the network [33].

Biologically, this architecture allows the learning subsystem to scale dramatically without a commensurate increase in the number of distinct neuron types or the amount of design information encoded in the genome. The decoupling of size from complexity offers a compelling explanation for the rapid evolutionary expansion of the hominin cortex, which doubled in size compared to chimpanzees over only 6 million years [34]. This expansion was achieved by the increasing proliferation of cortical progenitor cells [35–38], essentially "printing" more copies of the same computational template, rather than inventing new types. Moreover, comparative single-cell genomics has revealed that progenitors evolve distinct biases towards generating specific cell types, providing a direct mechanism for evolution to tune regional and species-specific hyperparameters [39,40].

3. The Steering Subsystem: Scaling via Diversification

The "steering subsystem"--comprising the hypothalamus, brainstem, and pallidum–operates on a different logic. These regions form the physiological core of the vertebrate brain, responsible for maintaining vital autonomic functions (breathing, arousal) and regulating essential homeostatic drives (e.g. hunger, thirst, reproduction, sleep, temperature). Rather than being blank slates waiting to be templated by experience, these regions implement specific logic and control circuits whose adaptive functionality has been shaped directly by evolution.

A useful analogy for understanding the role of the steering subsystem is "business logic" in software engineering, referring to the complex, bespoke code written to handle highly specific rules and procedures (e.g. tax codes or compliance checks). In contrast to cortex, which works by finding compressed representations of the world (e.g. faces, edges, or phonemes), the steering subsystem handles arbitrary contingencies and edge cases that must be enumerated individually.

Consider the algorithmic complexity required to properly implement what seem like simple physiological reflexes. For example, swallowing, far from being a simple reflex arc [41], relies on a medullary central pattern generator (CPG) to orchestrate precise, sequential activation of dozens of head muscles. Moreover, the CPG must be dynamically informed by real-time sensory feedback on breathing rate–to avoid aspiration into the airway–along with oral and stomach contents, among many other inputs. Swallowing is just one of hundreds of stereotyped behavioral patterns that must be stored, executed, and regulated by the steering subsystem.

These are not learned behaviors, but instead have been selected and refined across evolutionary time. They must therefore be encoded in an animal's genome. Moreover, because each of these behaviors addresses a unique problem, requiring unique inputs, outputs, and dynamics, they require specialized hardware. To support this, the steering subsystem follows **superlinear power-law scaling**: the number of distinct cell types rises rapidly alongside neuron count (Figure 3). This is consistent with the architecture observed in the cellular atlases: these steering system regions are collections of hundreds of distinct "micro-circuits," each composed of unique cell types dedicated to solving specific problems and computations. Redundancy may be present, but only enough to support resilience to injury, achieve a strong signal-to-noise ratio, and manage metabolic constraints, rather than increase representational capacity. A continued favorable evolutionary pressure exists to spawn new cell types, because they enable the organism and its progeny to anticipate complex sensorimotor contingencies without the risks and temporal delay of trial-and-error learning. This would help explain the observation that even highly conserved regions like midbrain continue to evolve novel cell types, even in recent evolutionary time [42].

 doi:10.20944/preprints202602.0767.v1

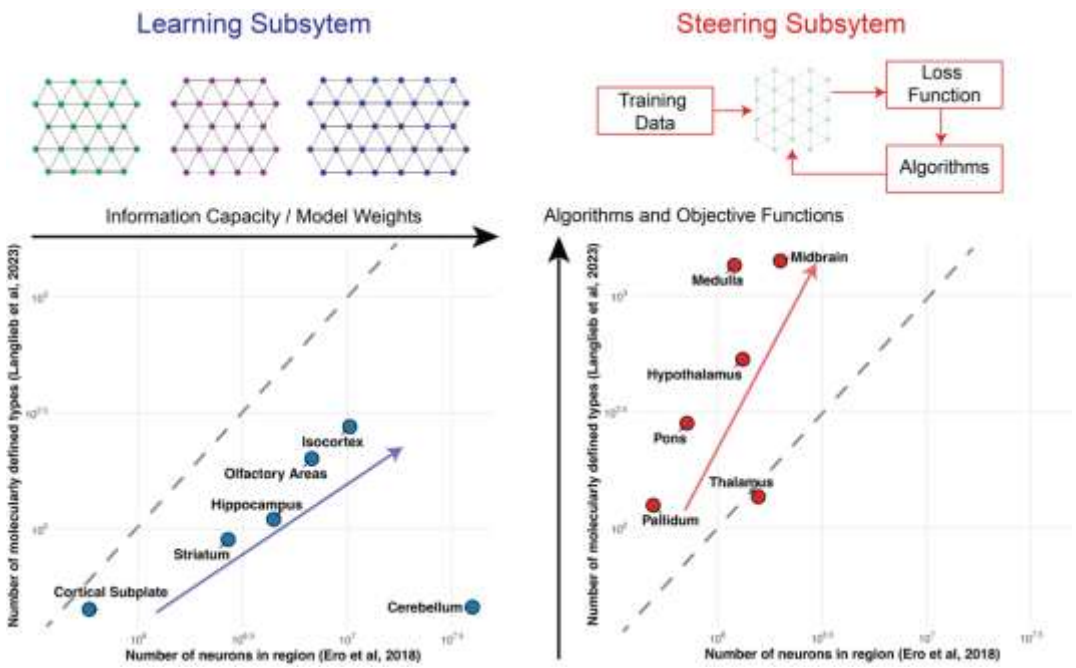


Figure 3. Scaling laws of the steering and learning subsystems. (Left) The learning subsystem represents the computational substrate of the brain for executing within lifetime learning algorithms. It exhibits a much lower power-law exponent between the number of neurons and number of cell types due to scaling through repeat architectural units (e.g. cortical minicolumn). (**Right**) The steering subsystem represents the set of algorithms and objective functions that are genetically encoded. It exhibits a much higher power law exponent (slope on log-log scale) between the number of neurons and the number of cell types, where the number of cell-types scales with the complexity of algorithms and objective functions.

The Steering Subsystem as a Biological Reward Function

The learning and steering subsystems interact in both directions. While the learning subsystem provides the competence to learn from the outside world, the steering subsystem provides the *goal*. In the framework of reinforcement learning (RL), the steering subsystem is the organism's reward function. For example, the reward function of the AlphaZero-chess engine [43] evaluates for checkmate, the win condition, and then returns a reward of +1 to the winner and –1 to the loser. All the intricate, superhuman chess strategies that the agent (the learning subsystem) displays ultimately flow from this "steering subsystem" module that defines what constitutes success or failure.

In biology, success involves navigating a complex ecological niche with many opportunities to fail, and no "do-overs." The complexity of survival necessitates a complex steering subsystem: a high-dimensional objective function that covers a vast array of contingencies. Consequently, humans, and all other animals, possess a long list of aversions, ranging from the physical–including pain, hunger, thirst, itch, nausea, and bitter tastes–to the emotional–such as loneliness, and embarrassment. Innate pleasures are equally wide-ranging and diverse. Remarkably, experimental evidence increasingly links specific preferences and aversions to specific cell types within the steering subsystem. In rodents, stimulation of specific cell groups can drive thirst [44], air hunger [45], itch [46], nausea [47], and even the experience of social isolation [48]. Moreover, extraordinary case studies of humans stimulated within steering subsystem components (the midbrain and hypothalamus) document subjects experiencing profound, complex emotions including despair [49], shame [50], extreme generosity [51], and déjà vu [52].

The steering system thus must serve as an inventory of these diverse drives, as well as a sophisticated arbiter, translating and prioritizing complex, disparate evolutionary fitness pressures into the immediate neural currency of pleasure, pain and drive. It is the "ghost in the machine" [53] that determines what the learning subsystem cares about.

 doi:10.20944/preprints202602.0767.v1

4. Comparisons with Existing Frameworks

To clarify our proposed "steering versus learning" architecture, it is useful to contrast it with existing dichotomies in neuroscience and psychology.

First, our "steering versus learning" is reminiscent of Paul MacLean's "triune brain" model [54] (reptilian vs. mammalian brain) or Jeff Hawkins's "old brain" and "new brain" [55]. We posit, however, that the most coherent dividing line between the two parts is computational rather than phylogenetic [56]. Both subsystems are evolutionarily ancient; the ancestor of all vertebrates possessed a pallium (learning) alongside a hypothalamus (steering), and even *drosophila* has a learning subsystem (the mushroom body). All of these learning subsystems—mushroom body, pallial cortical structures, and cerebellum—have maintained similar numbers of discrete cell types over evolutionary time. Furthermore, the distinction is not psychological: the steering subsystem is not just for "base" animal drives. Complex, "higher" motivations like compassion, social bonding, and curiosity also constitute innate drives encoded within the steering architecture.

Second, our "steering versus learning" bears some resemblance to the distinction between "nature versus nurture". However, the learning subsystem (nurture) is built upon algorithms and hyperparameters specified by the genome (nature). Conversely, the steering subsystem (nature) exhibits plasticity, changing the strengths of competing innate drives, and thus the organism's priorities, as an evolved response to different triggers like puberty [57], adiposity [58], or dominance [59]. Steering plasticity thus resembles the updating of variables in business logic, rather than the gradient descent-style weight updates of learning subsystem components like cortex.

By grounding the distinction in algorithmic function and scaling principles rather than historical or psychological categories, our framework provides a more mechanistic account of how cellular diversity maps onto behavioral complexity.

5. Maladaptation: When Fixed Drives Meet Changing Environments

If the steering subsystem acts as a reward function, it ultimately grounds the organism's sense of what is good and bad. This creates a vulnerability: if the reward function itself becomes maladaptive due to environmental changes, the learning subsystem will efficiently optimize on the wrong goals.

In biology, this vulnerability has been observed as instances of "phenological mismatch," where evolved timing mechanisms fall out of sync with the environment. For example, caribou time their seasonal migration using day lengths, while the plants they eat time their growth based on temperature. Climate change has desynchronized these signals, yet caribou continue to migrate by the calendar [60]. The caribou thus keep following their hardcoded innate drive, even as this drive becomes maladaptive.

Humans face a similar predicament. Our "habitat" has changed far more rapidly than our steering subsystem can evolve to accommodate. We possess innate drives to consume calorie-dense foods, conserve energy, and seek out social approval. In the modern environment, these drives are hijacked by novel stimuli engineered to activate these reward systems: processed sugar, addictive substances, sedentary entertainment, and social media. Evolutionary views of medicine in general [61,62], and of psychiatry specifically [63], regard the mismatch between ancestral drives and modern environments as main contributors to "diseases of modernity" such as obesity, addiction, anxiety and depression. The steering subsystem simply has not had time to update the "business logic" within the human genome.

In Artificial Intelligence (AI), this phenomenon is known as "reward hacking" or "specification gaming". Just as a human might short-circuit their evolutionary drive for nutrition by consuming refined sugar, an AI agent often finds unexpected, destructive shortcuts to maximize its digital reward signal without achieving the designer's actual intent. A famous example is an AI trained to play a boat racing game which, instead of winning the race, discovered it could achieve a higher score by endlessly spinning the boat in a circle to collect turbo-boost power-ups [64]. This parallel

 doi:10.20944/preprints202602.0767.v1

highlights a common challenge of any form of intelligence: a powerful "learning subsystem" will ruthlessly optimize its objective function. If that function–"the steering subsystem"--is misaligned with reality or the designer's true intent, the learning system will inevitably hack the reward, often with disastrous efficiency.

6. Architectural Vulnerabilities: Lessons for Neuropsychiatric and Neurodegenerative Disease

Current evidence suggests that neuropsychiatric illness likely operates at the interface between the learning and steering subsystems. On the one hand, cell type enrichment analyses of large-scale GWAS data have consistently placed the genetic liability for disorders like schizophrenia and bipolar disorder within the cellular hardware of the learning subsystem—specifically, the pyramidal neurons of the cortex and the medium spiny neurons (MSNs) of the striatum [25,65–67]. On the other hand, the pharmacological standard of care for these disorders—including antipsychotics, antidepressants, and mood stabilizers—largely targets signaling outputs of the steering subsystem, including dopamine, serotonin, and norepinephrine. These seemingly divergent pieces of evidence suggest that while the underlying structural vulnerability is often within the circuitry of the learning subsystem, the therapeutic lever involves adjusting the "steering" inputs that regulate it.

The recent clinical success of Glucagon-like peptide-1 (GLP-1) receptor agonists for treating addiction [68,69] reinforces the value of intervening at the interface between the two subsystems. Addictive disorders involve a highly heterogeneous constellation of symptoms in which affected patients develop a physiological dependence upon the drug, avoid normally rewarding activities or behaviors in favor of obtaining their drug of choice, and tolerate normally aversive experiences (e.g. social alienation) to sustain drug use. Studies have shown that GLP1R acts widely throughout the brain to suppress many of the individual behaviors that together make it difficult for a patient to stop using. Hundreds of cell populations in the steering subsystem express GLP1R, which together have been shown to modulate a wide range of behaviors, including aversion to drug use [70], drive towards relapse [71], and the immediate experience of reward when using the drug [72,73]. Meanwhile, GLP1R agonism in learning subsystem circuitry affects the experience of craving (e.g. in the nucleus accumbens of the striatum) [74], and contextual and environmental cues that drive use (e.g. in the ventral hippocampus and lateral septal outputs) [75]. This dual-system distribution activity enables GLP1R agonists to execute a coordinated restoration of equilibrium.

If GLP-1 agonists and classical neuromodulators are a guide, one path forward therapeutically lies in identifying additional signaling systems whose receptors bridge both subsystems. Fortunately, recent whole-brain cell atlases have revealed no shortage of such targets. For example, the cell atlases have identified dozens of neuropeptide signaling systems, often arising in steering cell types, that distribute the signal through specific receptors widely throughout the brain [25]. Systematically screening these understudied modulators could uncover new mechanisms to shift the learning-steering interface, offering therapeutic avenues for disorders that have resisted the traditional approach of monoaminergic modulation.

The divergence in cell scaling laws between the learning and steering subsystems also has important implications for how we conceptualize and detect neuronal vulnerability in neurodegenerative disease. Much of our current quantitative understanding of neuronal loss is derived from stereological counting in learning subsystem structures, such as the entorhinal cortex, hippocampus, and neocortex. In these regions containing many copies of each cell type, attaining statistical significance on cell loss counts in neurodegeneration is comparatively easy. For instance, in Alzheimer's disease, symptoms of hippocampal dysfunction are reliably detectable when the abundant pyramidal CA1 neurons typically exceed a loss threshold of approximately 20–30% [76]. However, the diversification scaling law of the steering subsystem – characterized by a long-tailed distribution where a minority of cell types are abundant but the vast majority are exceedingly rare [25,26]—renders traditional stereological approaches ineffective for most populations. Consequently, for a population of only a few thousand neurons, a symptomatic 30% loss might involve the

 doi:10.20944/preprints202602.0767.v1

disappearance of only a few hundred cells—a quantity that falls well within the noise floor of conventional counting methods. Compounding this statistical limitation, available measurements are predominantly cross-sectional and population-averaged, leaving the baseline variance of these rare cell types undefined. Consequently, genuine pathological attrition is often statistically indistinguishable from natural inter-individual heterogeneity, systematically masking vulnerabilities within the steering architecture.

Consistent with this limitation, historically, we have only robustly established selective loss in the steering subsystem when the cell populations are unusually distinct, such as when they are pigmented. The preferential vulnerability of dopaminergic neurons in the substantia nigra pars compacta is the pathological hallmark of Parkinson's disease, with motor symptoms emerging after approximately 30% of these pigmented neurons are lost [77]. Similarly, the visibly pigmented noradrenergic neurons of the locus coeruleus represent one of the earliest sites of tau pathology and neuronal loss in Alzheimer's disease [78]. It is likely no coincidence that the two most well-characterized vulnerable populations in the steering subsystem are those that are visibly distinct to the naked eye. It is thus possible that we detect degeneration in these populations not necessarily because they are uniquely vulnerable within the brain stem, but because their pigmentation and relative abundance make their absence easily quantifiable. Whole-brain cell atlases now provide the molecular parts list required to map rare populations; applying single-cell and especially spatial genomics tools systematically to disease tissue may reveal that neurodegeneration is far more driven by steering subsystem failure than previously appreciated.

7. Conclusion: Lessons of the Steering Subsystem for Neuroscience and AI Design

The revelation that the deep subcortical brain harbors the greatest diversity of neuronal cell types forces a re-evaluation of the underlying purpose of neural complexity. For decades, cellular neuroscientists have been searching the cortex and other "higher order" areas for distinct cellular specializations that make us human. But recent results suggest the opposite: the expansion of general intelligence–that is, improvements to the learning subsystem–has been mostly achieved through massive replication of a few scalable modules, tweaked for context-specific learning. By contrast, in the steering subsystem, evolution has taken on the task of encoding complex, high-dimensional survival logic with an expanding suite of unique cell types, tuned to wide-ranging aversions, preferences, and drives.

This contrastive framework illuminates a critical temporal asymmetry in biological intelligence. The learning subsystem allows for the rapid, "software-level" updating of behavior within a single lifetime. By contrast, the steering subsystem relies on "hardware-level" updates–the evolution of new cell types–which occur over far longer time scales. This disparity reframes the "diseases of modernity" as "phenological mismatches" where the hard-coded logic of the steering subsystem, tuned over millennia, fails to align with a rapidly changing environment. Recognizing that in fighting diseases like depression or addiction, we are fighting against the physical architecture of our own drives, underscores the importance of molecularly and cellularly unraveling the steering system's complexity. Developing novel strategies for the modulation of this system may be the missing key to unlocking treatments for mental illness, addiction, and personality disorders.

In artificial intelligence, this framework highlights a fundamental architectural gap. Contemporary models, including large language models, closely mirror the scaling laws of the learning subsystem: performance improves primarily through increases in parameter count, training data, and optimization time. Alignment, however, is typically imposed *post hoc* through reinforcement learning from human feedback (RLHF), in which a reward model is inferred from human preference judgments and used to fine-tune an already-trained network [79,80]. Consistent with this post-hoc nature, RLHF often reduces performance on likelihood-based and some task-specific benchmarks, even as it substantially improves human-rated usefulness and safety—highlighting a tradeoff between behavioral alignment and raw model capability [81]. Evolution, by

 doi:10.20944/preprints202602.0767.v1

contrast, implemented an elaborate, complex checks-and-balances system composed of many opposing drives to help organisms achieve their survival objectives. Consequently, the "alignment problem" may be a problem of architectural under-specification: we are attempting to align powerful learning models without the high-dimensional, evolutionarily grounded objective functions that biological brains possess [82–84].

Ultimately, the study of the steering subsystem remains the "dark matter" of the brain: vast and still largely unmapped, it is a hidden, powerful force that binds our behaviors into a coherent whole. That it scales through cell type diversification, rather than modular replication, mandates a different research strategy. We cannot necessarily rely on shared circuit principles, as in the study of cortical or cerebellar microcircuits; instead, we must commit to exhaustively characterizing the bespoke "business logic" encoded within thousands of unique cell types. Decoding how this immense diversity coalesces into stable objective functions promises to unlock new frontiers in medicine and philosophy alike, transforming how we treat neurodegenerative and psychiatric disease, how we understand the biological roots of moral intuition, and how we engineer safe and aligned artificial intelligence.

Author Contributions: Both authors contributed to conceptualization, writing, and revising this manuscript.

Acknowledgments: The authors would like to thank Steven Byrnes for generous and extensive discussions, the inspiration for the subsystem concept, and help during the writing process. We also thank Adam Marblestone for helpful conversations. This work was supported by the National Institutes of Health/National Institute of Mental Health (BRAIN RF1MH124598 and UM1MH130966) and the Stanley Center for Psychiatric Research. .

Competing interests: All authors declare no competing interests.

References

1. Ramón y Cajal, S. (1909). Histologie du système nerveux de l'homme & des vertébrés (Maloine).
2. Masland, R.H. (2001). The fundamental plan of the retina. Nat. Neurosci. *4*, 877–886.
3. Masland, R.H., and Sanes, J.R. (2015). Retinal ganglion cell types: Current states and lessons for the brain. Annu. Rev. Neurosci. *in press*.
4. Kim, I.J., Zhang, Y., Yamagata, M., Meister, M., and Sanes, J.R. (2008). Molecular identification of a retinal cell type that responds to upward motion. Nature *452*, 478–482.
5. Zhang, Y., Kim, I.-J., Sanes, J.R., and Meister, M. (2012). The most numerous ganglion cell type of the mouse retina is a selective feature detector. Proc. Natl. Acad. Sci. U. S. A. *109*, E2391–E2398.
6. Münch, T.A., da Silveira, R.A., Siegert, S., Viney, T.J., Awatramani, G.B., and Roska, B. (2009). Approach sensitivity in the retina processed by a multifunctional neural circuit. Nat. Neurosci. *12*, 1308–1316.
7. Masland, R.H. (2004). Neuronal cell types. Curr. Biol. *14*, R497–R500.
8. Kawaguchi, Y., and Kubota, Y. (1997). GABAergic cell subtypes and their synaptic connections in rat frontal cortex. Cereb. Cortex *7*, 476–486.
9. Gupta, A., Wang, Y., and Markram, H. (2000). Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex. Science *287*, 273–278.
10. Petilla Interneuron Nomenclature, Group, Ascoli, G.A., Alonso-Nanclares, L., Anderson, S.A., Barrionuevo, G., Benavides-Piccione, R., Burkhalter, A., Buzsaki, G., Cauli, B., Defelipe, J., et al. (2008). Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. Nat. Rev. Neurosci. *9*, 557–568.
11. Dong, H.W., Petrovich, G.D., and Swanson, L.W. (2001). Topography of projections from amygdala to bed nuclei of the stria terminalis. Brain Res. Brain Res. Rev. *38*, 192–246.
12. Dong, H.-W., and Swanson, L.W. (2004). Organization of axonal projections from the anterolateral area of the bed nuclei of the stria terminalis. J. Comp. Neurol. *468*, 277–298.
13. Swanson, L.W. (2000). Cerebral hemisphere regulation of motivated behavior11Published on the World Wide Web on 2 November 2000. Brain Res. *886*, 113–164.

 doi:10.20944/preprints202602.0767.v1

14. Zeng, H., and Sanes, J.R. (2017). Neuronal cell-type classification: challenges, opportunities and the path forward. Nat. Rev. Neurosci. *18*, 530–546.
15. Nelson, S.B., Sugino, K., and Hempel, C.M. (2006). The problem of neuronal cell types: a physiological genomics approach. Trends Neurosci. *29*, 339–345.
16. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.
17. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.
18. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.
19. Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat. Methods. https://doi.org/10.1038/nmeth.4407.
20. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090.
21. Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science *362*. https://doi.org/10.1126/science.aau5324.
22. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467.
23. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.
24. Shi, H., He, Y., Zhou, Y., Huang, J., Maher, K., Wang, B., Tang, Z., Luo, S., Tan, P., Wu, M., et al. (2023). Spatial atlas of the mouse central nervous system at molecular resolution. Nature *622*, 552–561.
25. Langlieb, J., Sachdev, N.S., Balderrama, K.S., Nadaf, N.M., Raj, M., Murray, E., Webber, J.T., Vanderburg, C., Gazestani, V., Tward, D., et al. (2023). The molecular cytoarchitecture of the adult mouse brain. Nature *624*, 333–342.
26. Yao, Z., van Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M., et al. (2023). A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature *624*, 317–332.
27. Zhang, M., Pan, X., Jung, W., Halpern, A.R., Eichhorn, S.W., Lei, Z., Cohen, L., Smith, K.A., Tasic, B., Yao, Z., et al. (2023). Molecularly defined and spatially resolved cell atlas of the whole mouse brain. Nature *624*, 343–354.
28. Erö, C., Gewaltig, M.-O., Keller, D., and Markram, H. (2018). A cell atlas for the mouse brain. Front. Neuroinform. *12*, 84.
29. Campbell, J.N., Macosko, E.Z., Fenselau, H., Pers, T.H., Lyubetskaya, A., Tenen, D., Goldman, M., Verstegen, A.M.J., Resch, J.M., McCarroll, S.A., et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. Nat. Neurosci. *20*, 484–496.
30. Chen, R., Wu, X., Jiang, L., and Zhang, Y. (2017). Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. Cell Rep. *18*, 3227–3241.
31. Byrnes, S.J. (2025). Intro to brain-like-AGI safety. https://doi.org/10.31219/osf.io/fe36n_v2.
32. Marblestone, A.H., Wayne, G., and Kording, K.P. (2016). Toward an integration of deep learning and neuroscience. Front. Comput. Neurosci. *10*, 94.
33. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. arXiv [cs.LG].

 doi:10.20944/preprints202602.0767.v1

34. Mora-Bermúdez, F., Badsha, F., Kanton, S., Camp, J.G., Vernot, B., Köhler, K., Voigt, B., Okita, K., Maricic, T., He, Z., et al. (2016). Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife *5*. https://doi.org/10.7554/eLife.18683.
35. Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proc. Natl. Acad. Sci. U. S. A. *109 Suppl 1*, 10661–10668.
36. Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F.K., Peters, J., et al. (2015). Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science *347*, 1465–1470.
37. Hansen, D.V., Lui, J.H., Parker, P.R.L., and Kriegstein, A.R. (2010). Neurogenic radial glia in the outer subventricular zone of human neocortex. Nature *464*, 554–561.
38. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323.
39. Schmitz, M.T., Sandoval, K., Chen, C.P., Mostajo-Radji, M.A., Seeley, W.W., Nowakowski, T.J., Ye, C.J., Paredes, M.F., and Pollen, A.A. (2022). The development and evolution of inhibitory neurons in primate cerebrum. Nature *603*, 871–877.
40. Nascimento, M.A., Biagiotti, S., Herranz-Pérez, V., Santiago, S., Bueno, R., Ye, C.J., Abel, T.J., Zhang, Z., Rubio-Moll, J.S., Kriegstein, A.R., et al. (2024). Protracted neuronal recruitment in the temporal lobes of young children. Nature *626*, 1056–1065.
41. Jean, A. (2001). Brain stem control of swallowing: neuronal network and cellular mechanisms. Physiol. Rev. *81*, 929–969.
42. Kamath, T., Abdulraouf, A., Burris, S.J., Langlieb, J., Gazestani, V., Nadaf, N.M., Balderrama, K., Vanderburg, C., and Macosko, E.Z. (2022). Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. Nat. Neurosci. *25*, 588–595.
43. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv [cs.AI].
44. Oka, Y., Ye, M., and Zuker, C.S. (2015). Thirst driving and suppressing signals encoded by distinct neural populations in the brain. Nature *520*, 349–352.
45. Chang, R.B., Strochlic, D.E., Williams, E.K., Umans, B.D., and Liberles, S.D. (2015). Vagal sensory neuron subtypes that differentially control breathing. Cell *161*, 622–633.
46. Dong, X., and Dong, X. (2018). Peripheral and central mechanisms of itch. Neuron *98*, 482–494.
47. Zhang, C., Kaye, J.A., Cai, Z., Wang, Y., Prescott, S.L., and Liberles, S.D. (2021). Area Postrema Cell Types that Mediate Nausea-Associated Behaviors. Neuron *109*, 461–472.e5.
48. Matthews, G.A., Nieh, E.H., Vander Weele, C.M., Halbert, S.A., Pradhan, R.V., Yosafat, A.S., Glober, G.F., Izadmehr, E.M., Thomas, R.E., Lacy, G.D., et al. (2016). Dorsal Raphe Dopamine Neurons Represent the Experience of Social Isolation. Cell *164*, 617–631.
49. Bejjani, B.P., Damier, P., Arnulf, I., Thivard, L., Bonnet, A.M., Dormont, D., Cornu, P., Pidoux, B., Samson, Y., and Agid, Y. (1999). Transient acute depression induced by high-frequency deep-brain stimulation. N. Engl. J. Med. *340*, 1476–1480.
50. Parvizi, J., Veit, M.J., Barbosa, D.A.N., Kucyi, A., Perry, C., Parker, J.J., Shivacharan, R.S., Chen, F., Yih, J., Gross, J.J., et al. (2022). Complex negative emotions induced by electrical stimulation of the human hypothalamus. Brain Stimul. *15*, 615–623.
51. Amstutz, D., Michelis, J.P., Debove, I., Maradan-Gachet, M.E., Lachenmayer, M.L., Muellner, J., Schwegler, K., and Krack, P. (2021). Reckless generosity, Parkinson's disease and dopamine: A case series and literature review. Mov. Disord. Clin. Pract. *8*, 469–473.
52. Hamani, C., McAndrews, M.P., Cohn, M., Oh, M., Zumsteg, D., Shapiro, C.M., Wennberg, R.A., and Lozano, A.M. (2008). Memory enhancement induced by hypothalamic/fornix deep brain stimulation. Ann. Neurol. *63*, 119–123.
53. Koestler, A. (1989). The ghost in the machine (Arkana).

 doi:10.20944/preprints202602.0767.v1

54. MacLean, P. (1973). A Triune concept of the brain and behaviour: Hincks memorial lectures T. J. Boag and D. Campbell, eds. (University of Toronto Press).
55. Hawkins, J., and Blakeslee, S. (2004). On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines (Times Books).
56. Cesario, J., Johnson, D.J., and Eisthen, H.L. (2020). Your brain is not an onion with a tiny reptile inside. Curr. Dir. Psychol. Sci. *29*, 255–260.
57. Naulé, L., Maione, L., and Kaiser, U.B. (2021). Puberty, A sensitive window of hypothalamic development and plasticity. Endocrinology *162*. https://doi.org/10.1210/endocr/bqaa209.
58. Pinto, S., Roseberry, A.G., Liu, H., Diano, S., Shanabrough, M., Cai, X., Friedman, J.M., and Horvath, T.L. (2004). Rapid rewiring of arcuate nucleus feeding circuits by leptin. Science *304*, 110–115.
59. Stagkourakis, S., Spigolon, G., Liu, G., and Anderson, D.J. (2020). Experience-dependent plasticity in an innate social behavior is mediated by hypothalamic LTP. Proc. Natl. Acad. Sci. U. S. A. *117*, 25789–25799.
60. Post, E., and Forchhammer, M.C. (2008). Climate change reduces reproductive success of an Arctic herbivore through trophic mismatch. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 2369–2375.
61. Williams, G.C., and Nesse, R.M. (1991). The dawn of Darwinian medicine. Q. Rev. Biol. *66*, 1–22.
62. Pennisi, E. (2011). Evolution. Darwinian medicine's drawn-out dawn. Science *334*, 1486–1487.
63. Stevens, A., and Price, J. (2021). Evolutionary Psychiatry (Routledge).
64. Faulty reward functions in the wild https://www.google.com/url?q=https://openai.com/index/faulty-reward-functions/&sa=D&source=docs&ust=1765851859210892&usg=AOvVaw2ZiaDdn7hBDDLY6bmNRRAK.
65. Mullins, N., Forstner, A.J., O'Connell, K.S., Coombes, B., Coleman, J.R.I., Qiao, Z., Als, T.D., Bigdeli, T.B., Børte, S., Bryois, J., et al. (2021). Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat. Genet. *53*, 817–829.
66. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.-Y., Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature *604*, 502–508.
67. Skene, N.G., Bryois, J., Bakken, T.E., Breen, G., Crowley, J.J., Gaspar, H.A., Giusti-Rodriguez, P., Hodge, R.D., Miller, J.A., Muñoz-Manchado, A.B., et al. (2018). Genetic identification of brain cell types underlying schizophrenia. Nat. Genet. *50*, 825–833.
68. Hendershot, C.S., Bremmer, M.P., Paladino, M.B., Kostantinis, G., Gilmore, T.A., Sullivan, N.R., Tow, A.C., Dermody, S.S., Prince, M.A., Jordan, R., et al. (2025). Once-weekly semaglutide in adults with alcohol use disorder: A randomized clinical trial. JAMA Psychiatry *82*, 395–405.
69. Klausen, M.K., Thomsen, M., Wortwein, G., and Fink-Jensen, A. (2022). The role of glucagon-like peptide 1 (GLP-1) in addictive disorders. Br. J. Pharmacol. *179*, 625–641.
70. Tuesta, L.M., Chen, Z., Duncan, A., Fowler, C.D., Ishikawa, M., Lee, B.R., Liu, X.-A., Lu, Q., Cameron, M., Hayes, M.R., et al. (2017). GLP-1 acts on habenular avoidance circuits to control nicotine intake. Nat. Neurosci. *20*, 708–716.
71. Hernandez, N.S., Weir, V.R., Ragnini, K., Merkel, R., Zhang, Y., Mace, K., Rich, M.T., Christopher Pierce, R., and Schmidt, H.D. (2021). GLP-1 receptor signaling in the laterodorsal tegmental nucleus attenuates cocaine seeking by activating GABAergic circuits that project to the VTA. Mol. Psychiatry *26*, 4394–4408.
72. Vallöf, D., Vestlund, J., and Jerlhag, E. (2019). Glucagon-like peptide-1 receptors within the nucleus of the solitary tract regulate alcohol-mediated behaviors in rodents. Neuropharmacology *149*, 124–132.
73. Schmidt, H.D., Mietlicki-Baase, E.G., Ige, K.Y., Maurer, J.J., Reiner, D.J., Zimmer, D.J., Van Nest, D.S., Guercio, L.A., Wimmer, M.E., Olivos, D.R., et al. (2016). Glucagon-like peptide-1 receptor activation in the ventral tegmental area decreases the reinforcing efficacy of cocaine. Neuropsychopharmacology *41*, 1917–1928.
74. Hernandez, N.S., O'Donovan, B., Ortinski, P.I., and Schmidt, H.D. (2019). Activation of glucagon-like peptide-1 receptors in the nucleus accumbens attenuates cocaine seeking in rats. Addict. Biol. *24*, 170–181.
75. Allingbjerg, M.-L., Hansen, S.N., Secher, A., and Thomsen, M. (2023). Glucagon-like peptide-1 receptors in nucleus accumbens, ventral hippocampus, and lateral septum reduce alcohol reinforcement in mice. Exp. Clin. Psychopharmacol. *31*, 612–620.

 doi:10.20944/preprints202602.0767.v1

76. La Joie, R., Perrotin, A., de La Sayette, V., Egret, S., Doeuvre, L., Belliard, S., Eustache, F., Desgranges, B., and Chételat, G. (2013). Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia. NeuroImage Clin. *3*, 155–162.
77. Greffard, S., Verny, M., Bonnet, A.-M., Beinis, J.-Y., Gallinari, C., Meaume, S., Piette, F., Hauw, J.-J., and Duyckaerts, C. (2006). Motor score of the Unified Parkinson Disease Rating Scale as a good predictor of Lewy body-associated neuronal loss in the substantia nigra. Arch. Neurol. *63*, 584–588.
78. Braak, H., Thal, D.R., Ghebremedhin, E., and Del Tredici, K. (2011). Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years. J. Neuropathol. Exp. Neurol. *70*, 960–969.
79. Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. arXiv [stat.ML].
80. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. arXiv [cs.CL].
81. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv [cs.CL].
82. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety. arXiv [cs.AI].
83. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. arXiv [cs.LG].
84. Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., and Irving, G. (2021). Alignment of language agents. arXiv [cs.AI].