THE EVOLUTION OF FUNDAMENTAL NEURAL CIRCUITS FOR COGNITION IN SILICO

By

Ali Tehrani-Saleh

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science – Doctor of Philosophy

2021

ABSTRACT

THE EVOLUTION OF FUNDAMENTAL NEURAL CIRCUITS FOR COGNITION IN SILICO

By

Ali Tehrani-Saleh

Despite decades of research on intelligence and fundamental components of cognition, we still know very little about the structure and functionality of nervous systems. Questions in cognition and intelligent behavior are addressed by scientists in the fields of behavioral biology, neuroscience, psychology, and computer science. Yet, it is difficult to reverse-engineer observed sophisticated intelligent behaviors in animals and even more difficult to understand their underlying mechanisms. In this dissertation, I use a recently-developed neuroevolution platform–called Markov brain networks–in which Darwinian selection is used to evolve both structure and functionality of digital brains. I use this platform to study some of the most fundamental cognitive neural circuits: 1) visual motion detection, 2) collision-avoidance based on visual motion cues, 3) sound localization, and 4) time perception. In particular, I investigate both the selective pressures and environmental conditions in the evolution of these cognitive components, as well as the circuitry and computations behind them. This dissertation lays the groundwork for an evolutionary agent-based method to study the neural circuits for cognition *in silico*.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Christoph Adami, for his help and support over my six years as his student. During my Ph.D., Chris was always available and willing to help me with my academic work and otherwise. I certainly would not have been this productive without Chris as my mentor. I also thank the past and current members of the Adami lab especially for helping me by sharing their knowledge and experience when I started as new a member of the Adami's lab. I also thank my committee members, Arend Hintze, Charles Ofria, J. Devin McAuley, and Wolfgang Banzhaf, for their guidance and comments during my Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 In search of building blocks of intelligence and cognition in light of artificial life

Scientists have long been studying animal behavior and brain function in search of components contributing to intelligent behavior, and how cognitive processes enable such behaviors that are essential to organism survival. These studies have taken a wide variety of approaches ranging from studying behaviors of animals in the wild to trained animals in the lab, and utilizing tools such as fMRI (Functional Magnetic Resonance Imaging) to genetic engineering in order to modify neural structure, and to building computational models to unravel mysteries of intelligence.

General intelligence has long been the holy grail of AI (Artificial Intelligence) and scientists have always been fascinated by the question: "can machines think?" [196]. But after decades of work and in spite of an exponential increase in computational power throughout this period, we still do not have a definitive answer. Researchers in the field of AI have constantly been speculating about possible routes that should be taken in order to advance toward or perhaps achieve general intelligence, yet controversies remain. It seems essential to me that our approach toward understanding intelligence, and perhaps building cognitive machines, must be through creating its building blocks first. I believe a bottom-up approach can take us closer to intelligence, by building up simpler and more fundamental cognitive widgets first and then attempt to join them together. As such, the main theme of this thesis is to build and study some of the simple yet fundamental neural circuits for cognition using neuroevolution. This approach enables us to study the components of cognition from an evolutionary standpoint where, I investigate the selective pressures and fitness landscape structures, as well as how they impact the evolved brains and their evolutionary history. This approach also allows us to analyze these cognitive components at different levels by investigating their behavioral characteristics, circuitry structure, and algorithms

and computations.

1.1.1 Why use computational evolution?

Ever since Charles Darwin published *On the Origin of Species*, his evolutionary theory has become the foundation of modern biology. In *the Descent of Man* he writes that the "mental faculties", similar to any other trait, vary in populations and are heritable, and as a result are subject to natural and sexual selection [40] (also see [18]). Thus, it seems inevitable to study intelligence and its building blocks in the light of evolution. Studying intelligence through computational evolutionary biology has been one of several active fields of research to shed light on intelligence alongside evolutionary psychology, evolutionary neuroscience, evolutionary behavioral ecology, etc. The advantage of using evolutionary methods in training artificial neural networks has started to attract more attention and is emphasized especially in recent years (see for example [208]). Scientists are slowly beginning to take advantage of neuroevolution because they are realizing that in order to build a simulated version of an intelligent organism it is only reasonable to follow the natural process by which intelligence has emerged in the first place, i.e., evolution. It should come as no surprise that we cannot reverse-engineer extremely complex biological brains, nor can we design machines with the same degree of complexity and performance. For example, Nguyen et al. used evolutionary computation to build images with random patterns that fooled CNNs (Convolutional Neural Networks) to classify them as actual images with high confidence [133]. It is noteworthy that image recognition is one of the leading areas in AI and scientists has been more successful in image processing compared to other areas such as natural language processing, social intelligence, or knowledge representation. This example and many similar findings [131, 80, 46, 178] underscores how far away a biological visual cortex is from our sophisticated designed image processing machines. This is perhaps an indication that we need to approach the problem from a higher level, for example by designing the substrate or components rather than designing the entire apparatus. Using evolutionary approaches enables us to avoid engineering the networks and let the evolutionary process take its course to build both the structure

and function of the network [49, 48, 51]. Furthermore, from an evolutionary perspective it may be more important to discover what selective pressures and environmental conditions might have resulted in the evolution of a particular intelligent behavior rather than understanding the behavior or the network. In other words, in an evolutionary process all we need is to build the right fitness landscape that leads to the evolution of the desired behavior. Ultimately, the problem of building general intelligence can be reduced to building the set of fitness landscapes within which we can evolve "thinking machines." Needless to say, building such fitness landscapes and evolving the agents within a proper substrate is still a very difficult problem and perhaps might as well be equally difficult as designing thinking machines from scratch.

Computational evolution has become of utmost interest to many scientists especially due to the rise of modern computers and the unprecedented increase in computational power. Computational evolution, and computational methods as a whole, are essential components of studying intelligence. Computational methods allow us to run "experiments" *in silico*, and easily change experimental parameters and explore conditions that have not been (or could not be) tested empirically. The computational models use different levels of abstraction to build the processing component, i.e. the brain, ranging from very detailed simulations of individual neurons, their networks, and their interactions with their environment such as neocortical column modeling [31] (NEURON platform) or Project Blue Brain [112], to less intricate models that partly capture the behavior of biological neurons such as common ANNs (Artificial Neural Networks) which are more efficient in computation and can achieve high performance in particular tasks such as pattern recognition.

1.1.2 Why Markov Brains?

Markov Brain Networks are a class of evolvable artificial brains in which populations of agents embedding *digital brains* undergo Darwinian evolution, through natural selection of inherited variations that increase an individual's ability to compete, survive, and reproduce. These digital brains have the Markov property, i.e., the future state of the network is influenced only by its present state. This property inspired the name Markov Network Brains, or Markov Brains for short.

More specifically, Markov Brains are neural networks in which neurons are binary variables that are connected via probabilistic or deterministic logic gates that represent synaptic excitatory or inhibitory connections. The connectivity and structure of the network, and the functionality of logic gates are determined by an evolutionary process. The aforementioned properties of Markov Brains makes them significantly distinct from other common artificial brain models such as ANNs. Some of these key differences are 1) evolvability of the network structure, 2) high variety in types of logic gates, 3) possibility of analysis of computations and algorithms of the evolved networks. In the following, I briefly describe these differences and their benefits and drawbacks and argue why using Markov Brains, of all artificial brain models, makes them suitable for the purpose of this thesis.

1. Evolvability of the network structure.

   Evolvability of the network structure is one of the key features that distinguishes Markov Brains from common ANNs. It is noteworthy that there are variations of ANNs that use evolution to train the network [205] and models such as NEAT (Neuroevolution of Augmenting Topologies) that enable the network structure to change during training [176]. However, researchers rarely use an evolutionary process or GA (Genetic Algorithm) to train ANNs and the evolutionary process is unnecessarily costly for ANNs because 1) ANNs are usually fully-connected networks that use real-valued numbers and as a result, require a lot of computation for each individual in the population, and 2) when training ANNs the structure of the network is almost always fixed, therefore, using a population of identical networks that are only different in their weights implies a lot of redundancy and is not computationally reasonable.

   On the other hand, Markov Brains are inherently sparse networks, which makes the computations required for the agents much cheaper, especially at the beginning of evolution. As the evolutionary process proceeds, the size of the networks grows, and their structure shapes to fit into the task, which is contrary to the conventional approach in training ANNs or CNNs

where researchers hand design the entire structure of the network while only the weights are subject to training. As mentioned before, one of the advantages of a top-down approach is that we do not inject our own biases into the engineering design of the network. Another advantage of this approach is that evolution provides us with a variety of network structures and functions (for example by evolving several populations) that perform the task (for example, see [34, 184, 185]) which then enables us to study the similarities and differences of a population of networks. It also has been shown that using an evolutionary approach in Markov Brains results in more sparse networks as opposed to ANNs that are fully connected networks [70]. This sparsity in connections has been shown to enable us to better detect and follow information in Markov Brains compared to ANNs [114]. For example, Marstaller et al. introduced an information-theoretic measure of "representation" and showed that Markov Brains evolved to perform an active categorical perception task have higher values of representation compared to ANNs evolved to perform the same task [116]. This is not to say that Markov Brains can evolve to have internal representation of the environment while ANNs cannot (note that any given Markov Brain can be recreated by a network of perceptrons that has the exact same computations and functions). Rather, the main differences in structure, connections, and functionality makes detecting and storing representations easier [70, 116].

2. High variety in types of logic gates.

   Markov Brains are networks of binary variables (neurons) that are connected via logic gates. These logic gates can take any number of inputs and based on their logic computation (logic table that is also subject to evolution) return a number of outputs. For example, a logic gate that takes two inputs and returns one output can have 16 different Boolean logic functions and as the number of inputs to a logic gate increase, the number of possible functions increases exponentially. This flexibility in functionality of logic gates in Markov Brains makes them more suitable especially for the purposes of this thesis. For example, it is well-known that single-layer perceptrons cannot perform an XOR (exclusive OR) operation [43] since the XOR operation is linearly inseparable. Thus, it is required to do the XOR operation in multiple

layers and with induction, as the non-linearity in the operation increases the required number of layers to perform it increases. On the contrary, Markov Brains handle such non-linearity in the operations in a more efficient way and as a result, they can evolve to be more sparse with higher information density.

Here I should mention that a recent method called Xnor-net has employed binary operations in convolution and filtering components of CNNs, and achieved state-of-art performance on the ImageNet dataset [158]. I should also acknowledge that an exponentially increased number of functions introduces an exponentially larger search space for optimization, but note that a set of smaller logic gates can always replace a larger set, and a smooth fitness landscape in which partially-optimized functions are rewarded is guaranteed to result in the optimum solution.

The more significant advantage of logic gates that connect neurons in Markov Brains is that they can mimic a more complex wiring in biological brains, with high density in synaptic or dendritic connections. For example, it has been shown that the non-linearity of dendritic connections makes them operate as computational subunits that take place before the summation at the synapse, which further facilitates pattern recognition in pyramidal neurons [150, 151]. Furthermore, Hawkins et al. show that a neuron with several thousand synapses segregated on active dendrites is capable of classifying several independent patterns and they can perform this task with large amounts of noise and variation introduced in those patterns [66]. Obviously, I am not suggesting that the logic tables of Markov Brains is an equivalent to more complex layered computations in dendritic and synaptic connections, but the more complex and non-linear computations of these logic tables and the accessibility of exponentially more complex functions is certainly in this respect, a closer model of biological neurons' connections compared to ANNs.

3. Possibility of analysis of computations and algorithms of the evolved networks.

   Understanding the mechanisms and algorithms at work in evolved networks is crucially

important for two main reasons. First, it seems necessary to understand the apparatus if we would want to correct its errors, prevent unexpected behavior, and improve its performance in the future. The second reason, which is more central to my thesis, is that we are attempting to recreate (evolve) biological-like brains in a machine with the purpose of discovering their structure and functionality, and then use this knowledge to better understand biological brains. This is also central to the entire field of Artificial Life, where the ultimate goal is to simulate living things *in silico* in order to discover out-of-reach mysteries, and to gain insight that helps us move forward in this journey.

As discussed earlier, while deep neural networks have been shown to be a powerful tool in AI, it is usually very difficult to understand the algorithms and computations behind their performance [206, 102, 80]. In other words, fpr the most part we have no clue as to how these huge networks that consist of components that perform sophisticated computations perform the desired task. It also seems impossible to translate their computations and algorithms to biological nervous systems and as a result, they cannot advance the task of understanding how an organism performs this specific task. In Markov Brains, on the other hand, there are methods that can reveal the mechanisms or algorithms that the agent utilizes in order to perform a particular task. Obviously it is not always very easy or straightforward to discover these computations, and it has been shown before that the evolved Markov Brain networks can be "epistemologically opaque" [116]. Yet, there are techniques that have been proposed and implemented that can unravel much about a Markov Brain's underlying mechanisms. For example, in chapters 2 and 4 I present an analysis of the types of computational components and their frequency distributions that are used in visual motion detection and sound localization tasks (also see [184, 182]). I also used knockout assays to measure how critical these computational components are in these evolved networks. In chapter 4, I performed transfer entropy measurements that can show the flow of information between neurons of the network. Furthermore, in chapter 5 I propose and use a technique based on the analysis of state-space transitions in Markov Brains when performing an event

duration judgment task (also see [185]). While the use of such techniques is still in its infancy, their ability to demonstrate important characteristics of the network and their algorithms was shown to be promising and points to their capacity to be enhanced in the future.

All said, the Markov Brains platform is a prominent substrate to study the evolution of intelligent behavior especially for the purpose of projects studied in this thesis. Furthermore, the Markov brains platform is recognized as one the recent specialized techniques in the neuroevolution community. For example, in a review on Evolutionary Algorithms, the authors categorize Markov brains as an innovation in the field of neuroevolution [173]. They write: "[Markov Brains] are showing some early promise especially in unsupervised learning" and attribute the success of the Markov brains to "being a more flexible substrate than ANNs, they could also lead to a more general understanding of the role recurrence plays in learning." The Markov Brains platform was used in numerous studies before and has been shown to be a powerful tool for the study of evolution of intelligence, such as evolution of predator-prey interactions [139, 140, 141, 137], active categorical perception [116, 142], image classification [34, 142], the evolution of neural plasticity [169, 170], the evolution of cognitive representations [45, 116, 89, 90, 91], the evolution of decision making strategies [97], and the dynamic interplay between ecology and brain structure [30, 141].

1.1.3 Cognitive Widgets: Fundamental Neural Circuits

The nervous system is undoubtedly the most complex organ/system in animals. For example, the human brain (which is a part of the central nervous system) consists of around 100 billion neurons (with about 20 billion in the neocortex alone) that differ in their anatomy, physiology, and functionality, with approximately 100 trillion connections. Our knowledge of the brain is still in its infancy, with numerous open questions and unknowns, given that the study of nervous systems, i.e., neuroscience, only dates back to Santiago Ramon y Cajal's seminal work in the 1890s. In fact, we still do not have a complete understanding of even much simpler central nervous systems like, for instance, an insect's brain, with only a few hundred thousand neurons. While there has been substantial research in neuroscience and its related fields, we have only come to understand

neural circuits and their functions for significantly simpler tasks. In particular, as we learn more about these systems, the more we realize the necessity of engaging experts from other disciplines and employing more specialized techniques for specific problems. Here, I focus on the following neural circuits and attempt to answer questions regarding their structure, functionality, and their evolutionary origins:

1. Visual motion detection.

   Visual motion detection is one of the fundamental components of visual perception and the computations take place at a low level (close to sensory neurons) in nervous system. Perceiving moving objects in the environment is crucial to an animal from an evolutionary point of view since it can be critical for survival; for example, detecting predators, prey, or falling objects [143]. One of the standard motion detection models was proposed by Werner Reichardt and Bernhard Hassenstein in the 1950s, based on a delay-and-compare scheme [65]. In addition to the Reichardt detector, researchers have proposed other types of motion detection models, such as edge-based models [113] and spatial-frequency-based models [6]. However, most computational motion detection models are based on the delay-and-compare scheme [143]. While motion detection in mammals and in particular humans is more complicated in structure and function, it is expected to have significant similarities to the basic Reichardt detector circuitry [22], and thus the Reichardt detector "module" is a key component of all motion detection circuits.

   In chapter 2 of this thesis, I study visual motion detection circuits and the underlying neuronal architectures. In particular, I study the distribution of different types of logic gates used to perform motion detection, the size of the network (the number of neurons contributing to the computation), and the presence of redundant logic gates, and their total complexity (i.e., number of logic gates). Furthermore, I investigate the evolutionary significance in complexity variation between circuits by seeding the population with a handwritten Reichardt detector circuit as the ancestor. I then ask whether an increase or decrease in their circuit complexity

is observed even though the performance of these circuits could not improve. If we observe a decrease in circuit complexity, it would suggest that the hand-written Reichardt detector could be further optimized, and therefore, we may be able to find simpler neuronal circuits in biological neuronal circuits. On the contrary, if we observe an increase in the complexity of the circuits, it would imply that other factors such as historical contingency or mutational robustness may be important factors in the evolution of visual motion detection circuits.

2. Intraspecific collision avoidance strategy based on apparent motion cues.

   The visual system is a significant perceptual component of an animal's cognitive system and provides it with information about its environment, for example when foraging for food, detecting predators or prey, and when searching for potential mates. Motion detection is one of the primary dimensions of visual systems [20] and plays a key role in decision making for most animals. In chapter 3 of this thesis, I study a specific type of behavior in *Drosophila melanogaster* (the common fruit fly), which is proposed as a collision-avoidance strategy based on visual motion cues. More precisely, I investigate the selective (i.e., evolutionary) pressures that might have given rise to this behavior.

   Fruit flies show an interesting behavior when perceiving two different types of optical flow in their retina, i.e., back-to-front and front-to-back motions. In a study published in 2009 by Branson et al., researchers investigated the walking trajectories of groups of fruit flies in a planar covered arena (so that they could only walk, not fly) using high-throughput recorded data to study the flies' behavior [25]. One of the results of their analysis showed that female fruit flies stop walking when they see another fly moving from back-to-front in their visual field (an optical flow referred to as "regressive motion") whereas they keep walking when they perceive conspecifics' motion from front-to-back in their visual field (referred to as "progressive motion"). Later, in a study published in 2012, Zabala et al. [207] further studied this behavior and tested a hypothesis that suggested that flies stop walking when perceiving regressive motion, and coined the term "regressive motion saliency". They used

a controllable fly-sized robot that interacted with a real fly in a planar arena. They used a robot instead of an actual fly in order to exclude other sensory cues such as image expansion ("looming," see [163]) and pheromones. Their results provided rigorous support for the regressive motion saliency hypothesis.

Subsequently, Chalupka et al. showed that a moving object (for example, another fly) that produces regressive motion in a fly's retina will reach the intersection point first whereas the fly that reaches the intersection first always perceives progressive motion on its retina [33]. They suggested a hypothesis called "generalized regressive motion" that suggests this behavior is a strategy to avoid collisions similar to the rules that ship captains use when moving on intersecting paths (see, e.g., [110]). However, it is not evident *a priori* which selective pressures or environmental circumstances could give rise to this behavior. For example, it is unclear whether collision avoidance alone could be a significant enough evolutionary factor for this behavior. In chapter 3, I test whether collision avoidance can be a sufficient selective pressure for the evolution of this behavior. I also investigate the environmental conditions, such as the varying costs and benefits involved, in the evolution of the described behavior. I also explore how the interplay (and trade-offs) between the necessity to move and the avoidance of collisions can result in the evolution of regressive motion saliency in digital flies.

3. Sound localization.

   Sound localization is another one of the fundamental cognitive neural circuits that has been widely studied [130, 149]. Sound localization mechanisms in mammalian auditory systems function based on various cues such as interaural time difference, interaural level difference, etc. [128]. Interaural time difference (which is the main cue behind the sound localization mechanism) is the difference between the times at which sound reaches the two ears. One of the most prominent sound localization models was proposed by Jeffress [79], in which sound reaches the right ear and left ear at two possibly different times. These stimuli are then

processed in a sequence of *delay* components and reach an array of detector neurons. Each detector fires only if the two signals from different pathways, the left ear pathway and the right ear pathway, arrive at that neuron simultaneously.

I used sound localization circuits as a benchmark to study how well transfer entropy analysis [165] can capture the information flow in neural circuits. Markov Brains have been shown to be a suitable platform to study the information-theoretic correlates of fitness and network structure in neural networks [45, 8, 164, 114, 85]. The Markov Brains platform enables us to analyze structure, function, and circuitry of hundreds of evolved neural circuits. As a result, I can perform statistical analysis on these evolved circuits (as opposed to studying only a single evolutionary outcome), for example, investigate the frequency of different types of relations, and further assess how crucial different operators are for each evolved task, by performing knockout experiments in order to measure an operator's contribution to the task.

4. Time perception.

   Time perception refers to the subjective experience of time that can be measured by someone's own perception of the duration of events. Time perception is a key component in our ability to deduce causation, to predict, infer, and forecast. As a result, time perception plays a key role in the survival of an organism by predicting and deciphering events in the world [134, 160]. The event duration perception is not objective, rather, we perceive temporal signals subjectively, and our perception is influenced by various factors such as attention [193, 37, 32, 108, 188]. A central hypothesis in time perception posits that the more attention devoted to the temporal characteristics of an event, the longer it is perceived [193, 37, 32, 108, 188]. There are several competing models of time perception. In models such as Scalar Expectancy Theory (SET) [54], it is assumed that event duration perception is performed with computations similar to that of an internal clock [53, 54, 191]. Models like SET also assume that in such an internal clock, the amount of attention allocated to the stimulus is adjusted based on the amount of attention, and that the attention is uniformly distributed in time. On the

contrary, in models such as Dynamic Attending Theory (DAT) [81, 82, 101] the attention is not distributed uniformly in time, rather, the temporal structure of the stimulus may increase or decrease levels of attention in time. In particular, rhythmic stimuli *entrain* the brain and lead to periodic peaks and troughs of attention.

Interval timing models such as DAT and SET and their computational counterparts usually take a top-down approach, meaning they are designed based on a set of rules so that they can describe behavioral/psychophysical data in duration perception [53, 81, 82, 44, 117]. In chapter 5 of this thesis, I take a bottom-up approach where I evolve a population of artificial brains consisting of lower-level components. In particular, I use Darwinian evolution to create artificial digital brains that are able to perform duration judgments in auditory oddball paradigms, similar to experiments performed by Fromboluti and McAuley [121]. I then study the evolved brains as though they are participants in a psychophysical experiment. For example, I investigate psychometric parameters of the evolved brains, such as their discrimination thresholds. I also can test these brains when exposed to stimuli patterns that they have not experienced during evolution such as arrhythmic stimuli. Furthermore, I can investigate the algorithms and computations involved in duration judgment, and analyze how these algorithms allocate attention to different parts of the stimuli. This analysis can demonstrate the similarities and differences of the evolved duration perception mechanisms and the underlying mechanisms of SET and DAT models.

In this thesis, I studied only a few cognitive circuits, while there are many other well-studied visual and auditory neural circuits as well as cognitive components. In the future, it is imperative that we also investigate other modes of sensation such as olfaction and touch.

1.2 Outline

In the following chapters, I address various questions concerning neuronal circuits and present my findings. In Chapter 2 I show how the evolution of motion detection circuits in Markov brains can lead to a diversity of circuits with a variety of structures in gate compositions and with different

levels of complexity. I also show that the complexity variation in evolved brains circuitry is due to selection for mutational robustness. These results suggest that different species may evolve different circuits for similar neuronal functions. In Chapter 3 I present the evolution of collision avoidance in digital flies, test the "generalized regressive motion" hypothesis, and discuss the environmental conditions and selective pressures that could give rise to this behavior. In Chapter 4 I investigate whether transfer entropy measurements can infer the information flow in two different neuronal circuits: visual motion detection and sound localization. In Chapter 5 I present the evolution of Markov brains that solve the event duration judgment task, and how the analysis of the underlying algorithms performed by digital brains can challenge existing models of time perception. In Chapter 6 I present a summary of my findings in completed projects and then I propose possible directions for future research. My proposal is the evolution of Markov Brains that perform image classification via saccadic eye movements.

In the remainder of this section, I briefly explain the findings from my finished projects that are presented in more detail in Chapters 2-5.

1.2.1 Visual motion detection

A central goal of evolutionary biology is to explain the origins and distribution of diversity across life. Beyond species or genetic diversity, we also observe diversity in the circuits (genetic or otherwise) underlying complex functional traits. However, while the theory behind the origins and maintenance of genetic and species diversity has been studied for decades, theory concerning the origin of diverse functional circuits is still in its infancy. It is not known how many different circuit structures can implement any given function, which evolutionary factors lead to different circuits, and whether the evolution of a particular circuit was due to adaptive or non-adaptive processes. Here, I use digital experimental evolution to study the diversity of neural circuits that encode motion detection in digital (artificial) brains. I find that evolution leads to an enormous diversity of potential neural architectures encoding motion detection circuits, even for circuits encoding the exact same function. Evolved circuits vary in both redundancy and complexity (as previously found

in genetic circuits) suggesting that similar evolutionary principles underlie circuit formation using any substrate. I also show that a simple (designed) motion detection circuit that is optimally-adapted gains in complexity when evolved further, and that selection for mutational robustness led this gain in complexity.

1.2.2 Collision avoidance mechanisms in *Drosophila melanogaster*

Flies that walk in a covered planar arena on straight paths avoid colliding with each other, but which of the two flies stops is not random. High-throughput video observations, coupled with dedicated experiments with controlled robot flies have revealed that flies utilize the type of optic flow on their retina as a determinant of who should stop, a strategy also used by ship captains to determine which of two ships on a collision course should throw engines in reverse. I use digital evolution to test whether this strategy evolves when collision avoidance is the sole selective pressure. I find that the strategy does indeed evolve in a narrow range of cost/benefit ratios, for experiments in which the "regressive motion" cue is error free. I speculate that these stringent conditions may not be sufficient to evolve the strategy in real flies, pointing perhaps to auxiliary costs and benefits not modeled in our study.

1.2.3 Information flow in evolved *in silico* motion detection and sound localization circuits

How cognitive neural systems process information is largely unknown, in part because of how difficult it is to accurately follow the flow of information from sensors via neurons to actuators. Measuring the flow of information is different from measuring correlations between firing neurons, for which several measures are available, foremost among them the Shannon information, which is an undirected measure. Several information-theoretic notions of "directed information" have been used to successfully detect the flow of information in some systems, in particular in the neuroscience community. However, recent work has shown that directed information measures such as transfer entropy can sometimes inadequately estimate information flow, or even fail to identify manifest directed influences, especially if neurons contribute in a cryptographic manner

to influence the effector neuron. Because it is unclear how often such cryptic influences emerge in cognitive systems, the usefulness of transfer entropy measures to reconstruct information flow is unknown. Here, I test how often cryptographic logic emerges in an evolutionary process that generates artificial neural circuits for two fundamental cognitive tasks (motion detection and sound localization). Besides counting the frequency of problematic logic gates, I also test whether transfer entropy applied to an activity time-series recorded from behaving digital brains can infer information flow, compared to a ground-truth model of direct influence constructed from connectivity and circuit logic. These findings suggest that transfer entropy will sometimes fail to infer directed information when it exists, and sometimes suggest a causal connection when there is none. However, the extent of incorrect inference strongly depends on the cognitive task considered. These results emphasize the importance of understanding the fundamental logic processes that contribute to information flow in cognitive processing, and quantifying their relevance in any given nervous system.

1.2.4 Evolution of event duration perception and implications on attentional entrainment

While cognitive theory has advanced several candidate frameworks to explain attentional entrainment, the neural basis for the temporal allocation of attention is unknown. Here I present a new model of attentional entrainment guided by empirical evidence obtained using a cohort of 50 artificial brains. These brains were evolved *in silico* to perform a duration judgment task similar to one where human subjects perform duration judgments in auditory oddball paradigms. I found that the artificial brains display psychometric characteristics remarkably similar to those of human listeners, and exhibit similar patterns of distortions of perception when presented with out-of-rhythm oddballs. A detailed analysis of mechanisms behind the duration distortion suggests that attention peaks at the end of the tone, which is inconsistent with previous attentional entrainment models. Instead, the suggested model of entrainment emphasizes increased attention to those aspects of the stimulus that the brain expects to be highly informative.

CHAPTER 2

EVOLUTION LEADS TO A DIVERSITY OF MOTION-DETECTION NEURONAL CIRCUITS

One of the most astonishing aspects of life is the overwhelming amount of diversity that has existed throughout life's history. Ever since Charles Darwin published *On the Origin of Species*, evolutionary biologists have tried to understand the processes that lead to biological diversity [41]. On the micro scale, the question of how genetic diversity is maintained within a population has been of interest to population geneticists [89, 105, 180, 12] for decades; work on this topic still continues to this day [58]. In a similar fashion, ecologists have long been interested in the ecological and evolutionary processes that lead to the origins [156, 135] and maintenance [162, 35] of species diversity. The rise of cheap sequencing technologies in recent years has led to the recognition of another characteristic of biological diversity, molecular diversity [187], or diversity in the sense that multiple genotypes can lead to the same phenotype [57]. In other words, evolution can lead to a diversity of genetic circuits across species [194].

The evolutionary principles that lead to molecular diversity in genetic systems has been well-explored. The relationship between genotype and phenotype must be many-to-one to allow for the existence of neutral evolutionary trajectories between genotypes. Computational studies of metabolic networks, gene regulatory networks, and RNA-structure networks [reviewed in [200]] all show evidence of neutral paths that conserve phenotypes between different genotypes. Many-to-one genotype-phenotype mappings are even present in artificial digital evolution systems [e.g., [100, 50, 99]] and evolutionary simulations of digital logic circuits [157]. Empirical studies of biological systems suggest the existence of multiple genotypes encoding similar phenotypes, either through genetic analysis [195, 181], comparative genomics [39], or experimental evolution [106, 73]. However, the evolutionary reasons why populations evolve one genotype instead of another genotype, or which evolutionary processes lead to the evolution of different genotypes, are largely unexplored in biological systems due to the difficulty of deciphering every possible evolutionary

trajectory and process, and the waiting time required for many of these evolutionary events to occur [but see [106]]. This difficulty presents a prime opportunity for artificial life and digital evolution studies to perform “digital genetics" and test hypotheses for why some populations, but not others, evolve certain genotypic characteristics [1].

Genetic circuits are not the only biological network shaped by evolution. Neuronal circuits are also shaped by selective pressures, and much work has been devoted to understand those. Much of the literature, however, has focused on whether evolution optimizes the wiring patterns of a brain, or the efficiency of the circuitry [see, e.g., the discussion in chapter 7 of [175]]. For example, it is clear that the wiring pattern of the neuronal circuitry of the roundworm *Caenorhabditis elegans* is not optimal [7]. At the same time, there appear to be certain network motifs that are strongly favored in the worm brain [154], suggesting that evolution has a hand in optimizing computational efficiency. However, very little is known about the wiring diversity underlying circuits with the *same* function. According to the principles of evolvability and robustness discussed above, such diversity could be key for the adaptability of brains. In fact, both modeling [152] and empirical [56] studies have shown that neuronal circuits can vary in their internal parameters but lead to the same functional output [111]. And while many of these studies examine variation within one species [56], similar results have also been found between species, suggesting evolutionary mechanisms can also cause these differences [171]. This outcome is not surprising, as evolution and natural selection is expected to primarily act on the function, not the circuit encoding said function [171]. These results motivate the question as to how and why evolution leads to neuronal circuits with different characteristics for the same function.

Here we use digital evolution to study the evolution of neuronal circuits for visual motion detection. Perception of moving objects in the environment is of utmost significance from an evolutionary standpoint since it can be critical to survival of animals (including humans); detecting predators, prey, or falling objects can pose a *live or die* question [143]. In the 1950s, Werner Reichardt along with Bernhard Hassenstein proposed a simple computational model [now known as the Reichardt detector], that is based on a delay-and-compare scheme [65]. The main idea behind

this model is that a moving object stimulates two adjacent receptors (or regions) in the retina at two different time points. In Fig. 2.1(A), an object (a star) is moving from left to right stimulating two adjacent receptors n1 and n2, at time points t and $t + \Delta t$. In the neural circuit illustrated in Fig. 2.1(A), which is a portion of the entire Reichardt detector circuit, τ functions as a temporal filter that *delays* the received stimulus from receptor n1. This delayed signal will then be multiplied (in the $\times$ neuron) with the stimulus received in n2 at $t + \Delta t$. This multiplication result, therefore, detects motion from left to right. However, this half-circuit only detects motion in one direction. In the full Reichardt detector circuit shown in Fig. 2.1(B), the outcome of the multiplication from two similar computations, but in opposite directions, are subtracted. Thus, the result will be a positive value for left to right motion (also called *preferred direction*, PD), and negative for right to left motion, (termed the *null direction*, ND).

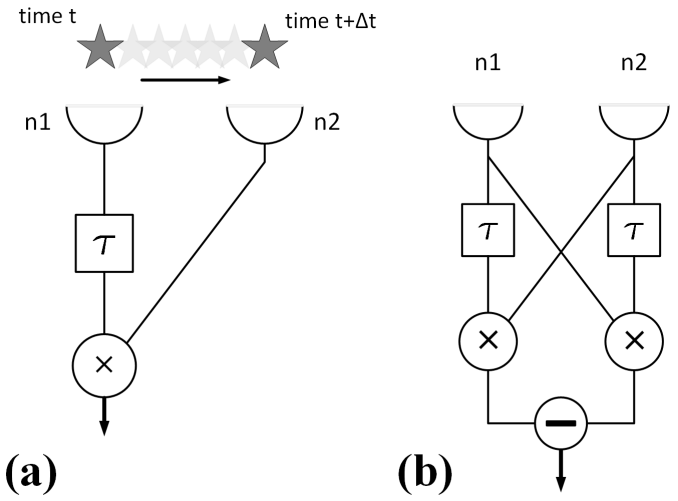

Figure 2.1: (A) A half Reichardt detector circuit. An object (star) moving from left to right stimulating two adjacent receptors, n1 and n2, at time points t and $t + \Delta t$. (B) A full Reichardt detector circuit. In full Reichardt detector circuits, the results of the multiplications from each half circuit are subtracted.

Beyond the Reichardt detector, other types of motion detection models were also proposed, e.g. edge-based models [113] and spatial-frequency-based models [6]. However, most computational motion detection models are based on the delay-and-compare scheme [143]. For example, the Barlow-Levick (BL) motion detection model [13] is similar to the Reichardt model in that it also

employs asymmetric temporal filtering of signals that are then fed to a non-linearity component, but they differ in the location of the filter and type of non-linearity component. While motion detection in mammals and in particular humans is expected to be far more complex, there are significant similarities to the basic Reichardt detector logic [22], and thus the Reichardt detector “module" of motion detection is likely a key component of all motion detection circuits.

Using digital experimental evolution methods, we found that motion detection circuits can be encoded by a wide diversity of neuronal architectures. Evolved brains differ in the logic gates used to perform motion detection, in the wiring between these logic gates, in the presence of redundant logic gates, and in their total complexity (i.e., number of logic gates). We explored the evolutionary significance in complexity variation between brains by evolving brains using a handwritten optimal motion detection circuit as the ancestor. These brains also increased in complexity although no improvement in the performance of their circuit could occur. Instead, these brains evolved greater complexity due to selection for mutational robustness. These results suggest that different species may evolve different circuits for similar neuronal functions.

2.1 Methods

In this study, we use an agent-based model to study evolution of computational visual motion detection circuits. In this model, agents embody neural networks known as “Markov brains" (MB) [69]. Markov brains have three different types of neurons that help the agent interact with the outside world: 1) sensory neurons, that receive the information from the environment, 2) hidden neurons that assimilate the agent’s processing unit, and 3) decision (“motor") neurons that function as the actuators of the agent. In other words, sensory neurons are written to by the surrounding environment, hidden neurons process the received information, and the decision neurons specify the actions of the agent in its environment.

Markov brains are evolvable networks of neurons in which the neurons are connected via probabilistic/deterministic logic gates. In the experimental setup used in this study, a logic neuron

is a binary variable whose state is either 0 or 1 (it is quiescent or it fires[1]). The states of the neurons are updated in a Markov fashion, i.e., the probability distribution of states of the neurons at time step $t+1$ depends only on the states of neurons at time step t as shown in Fig. 2.2(A). That figure shows a Markov brain with 11 neurons and two hidden Markov gates (HMG) at two consecutive time steps t and $t+1$. Hidden Markov gates determine how the states of the neurons at time step t+1 are updated given the states of the neurons at time t. For example in Fig. 2.2(B), gate 1 takes the states of neurons 0, 2, and 6 as inputs and writes updated states to output neurons 6 and 7. Each hidden Markov gate has a probabilistic logic table that specifies the probability of every possible output given the states of the input (Fig. 2.2(C)). That figure shows the probability table of gate 1 with 8 rows for all possible input states, and 4 columns for each possible output states (note that there are $2^3 = 8$ possible input states for 3 binary inputs, and similarly, $2^2 = 4$ for outputs). Each entry in the table represents the probability of a specific output, given a particular input. For instance, p_{53} represents the probability of getting output states $\langle 1, 1\rangle$, with decimal representation 3, given the input states $\langle 1, 0, 1\rangle$, with decimal representation 5. As a result, the sum of the probabilities of each row should be equal to 1. In this work, we constrain hidden Markov gates to be deterministic, therefore, the output states will always be the same given a particular input (probabilities in the table are either 0 or 1 and only one entry in each row of the table can be equal to 1). Markov brains can evolve to perform a variety of tasks such as active categorical perception [115], swarming in predator-prey interactions [139], collision avoidance strategies using optical flow classification in fruit flies [186], and decision making strategies in humans [97]. In the evolutionary process, the connections of the networks and the underlying logic of the connected gates change (evolve), and therefore, the agents adapt to their environment. More specifically, the number of gates, how each gate is connected to its inputs/outputs neurons, and the logic table of the gates are subject to evolution. However, the total number of neurons, the number of each type of neurons (i.e., sensory neurons, hidden neurons, and decision neurons), does not change during evolution. In our experimental setup for instance, we use MBs with 16 neurons in which two neurons (neurons 1

[1]These logic neurons are thought to represent the state of groups of biological neurons.

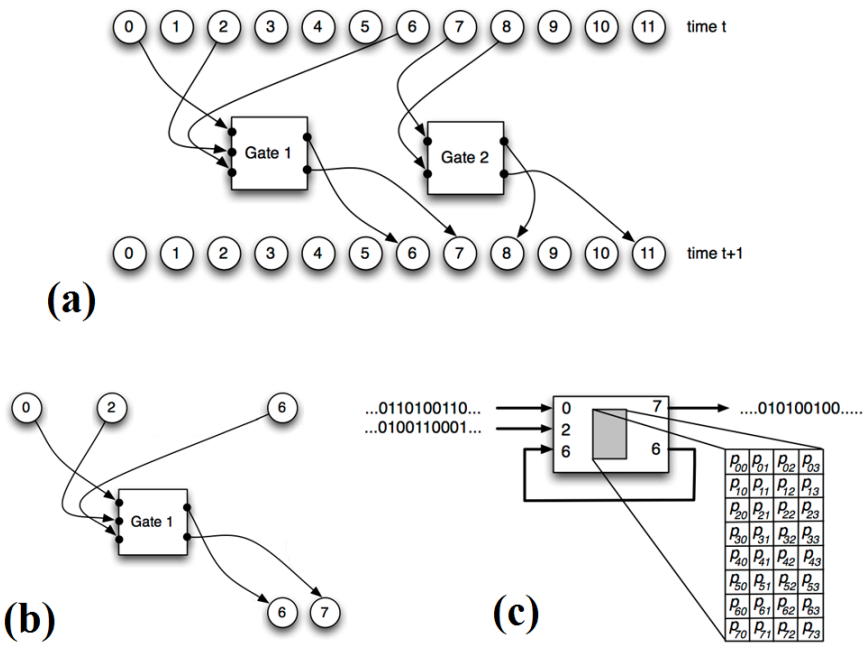

Figure 2.2: (A) A Markov brain with 11 neurons and 2 gates shown at two time steps t and $t + 1$. The states of neurons at time t and the logic operations of gates determine the states of neurons at time $t + 1$. (B) One of the gates of the MB whose inputs are neurons 0, 2, and 6 and its outputs are neurons 6 and 7. (C) Probabilistic logic table of gate 1.

and 2) are designated as sensory neurons, and two neurons (neurons 15 and 16) are assigned as decision neurons, while the remaining 12 neurons are hidden neurons. In order to evolve MBs, we

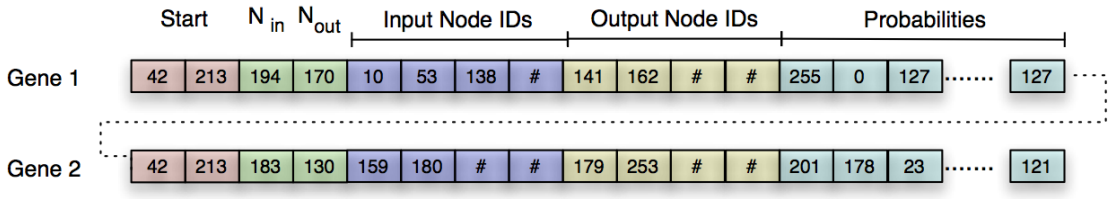

Figure 2.3: A Markov brain is encoded in a sequence of bytes that serves as the agent's genome.

apply a Genetic Algorithm (GA) to a population of MBs in which each MB is encoded in a genome as shown in Fig. 2.3. The genome of each MB is a sequence of numbers in the range [0,255] (a sequence of bytes) that encodes hidden Markov gates (HMGs), their connections, and their logic. The arbitrary pair $\langle 42, 213 \rangle$ is chosen as the start codon for each gate. The next two bytes following

the start codon encode the number of inputs and the number of outputs of the HMG, respectively. In our experimental setup, we constrained MBs to always have 2 inputs and 1 output, therefore, these two bytes are ignored in transcription. The subsequent (downstream) loci in the genome encode which neurons are connected to this HMG as input, which neuron is connected to the output, and finally the logic table of the HMG.

In our experimental setup, we initialized the populations with 100 genomes with 5,000 random bytes. We sprinkled those random bytes with four start codons in each genome to speed up initial evolution. Thus, all genomes in the initial population have at least four random HMGs. As mentioned before, all HMGs in our setup are deterministic and have 2 inputs and 1 output. As a result, HMGs can only have 16 possible logic tables. We ran 100 replicates of this experiment for 10,000 generations with mutations, roulette wheel selection, and 5% elitism. The GA configuration is presented in more detail in Table 5.2.

Table 2.1: Genetic Algorithm configuration. We evolved 100 populations of 100 MBs for 10,000 generations with point mutations, deletions, and insertion. We used roulette wheel selection, with 5% elitism, and with no cross-over or immigration.

Population size	100
Generations	2000
Initial genome length	5,000
Initial start codons	4
Point mutation rate	0.5%
Gene deletion rate	2%
Gene duplication rate	5%
Elitism	5%
Crossover	None
Immigration	None

The fitness function is designed in order to evolve MBs that function as a visual motion detection circuit. In doing so, two sets of stimuli are presented to the agent in two consecutive time steps and the agent classifies the input as either: motion in preferred direction (PD), stationary, or motion in null direction (ND). Neurons 1 and 2 (the sensory neurons) represent two adjacent receptors separated by a fixed distance that can sense the presence or the absence of a visual stimulus. The binary value of the sensory neuron becomes 1 when a stimulus is present, and it becomes

(or remains) 0 otherwise (see Fig. 2.4). Thus, there are 16 possible sensory patterns that can be presented to the agent (2 binary neurons at 2 time steps). Among these 16 input patterns, 3 input patterns are PD, 3 are ND, and the other 10 are stationary patterns. Agents classify the sensory pattern with 2 decision neurons, neurons 15 and 16. We assigned the sum of the values of the decision neurons to represent the category of the sensory pattern: when both decision neurons fire (sum=2), the sensory pattern is classified as PD, when only one of the decision neurons fires (sum=1), the sensory pattern is classified as stationary, and when neither fire the sensory pattern is classified as ND (sum=0). We chose this encoding for three classes of input pattern to facilitate the evolution of motion detection circuits. In preliminary experiments, we tried three different methods of encoding input pattern classes and found this one to evolve the fastest. In those preliminary experiments, we tried the following alternative encodings: assigning one neuron to each class (i.e. three decision neurons), assigning the decimal value of the pair of decision neurons to each class, i.e., 00 $\rightarrow$ ND, 01 $\rightarrow$ stationary, 10 $\rightarrow$ PD, and ignore 11, and finally assigning the sum of the values of decision neurons to each class. For the last two encoding methods, we tried all possible permutations of encodings and the one we chose consistently leads to the best results.

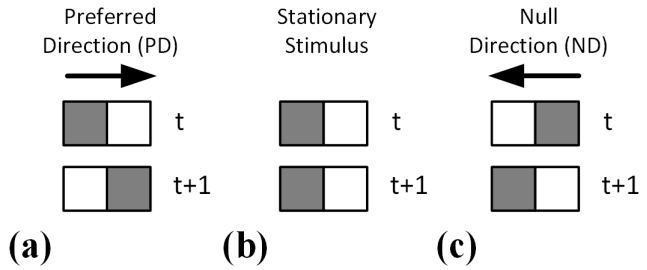

Figure 2.4: Schematic examples of three types of input patterns received by the two sensory neurons at two consecutive time steps. Grey squares show presence of the stimuli in those neurons. (A) Preferred direction (PD). (B) Stationary stimulus. (C) Null direction (ND).

All agents of the population are evaluated in all 16 possible sensory patterns and gain a reward for correct classification (no reward or penalty for incorrect classifications). The reward values for correct classifications of each class is inversely proportional to their frequency: the reward for PD and ND patterns are 10, and the reward value for correct classification of stationary patterns are

3. However, in the results presented in the next section, all fitness values are normalized to take a maximum value of 100.

2.2 Results

After evolving 100 populations for 10,000 generations, we isolated one of the genotypes with the highest score from each population and analyzed its ability to perform the same function as a motion detection circuit. Seventy-five of the one hundred brains evolved a perfect motion detection circuit (correct classification of all 16 patterns); we used those brains for the rest of our analysis. A preliminary analysis of our evolved brains suggested that evolution led to a wide diversity of neuronal circuit architectures. Amongst our population of 75 brains, we found both relatively simple neuronal circuits (Fig. 2.5(A)) and more complex neuronal circuits (Fig. 2.5(B)), suggesting that not only does evolution lead to a large number of different motion detectors, but they also all vary in complexity (defined here as the number of gates composing a circuit).

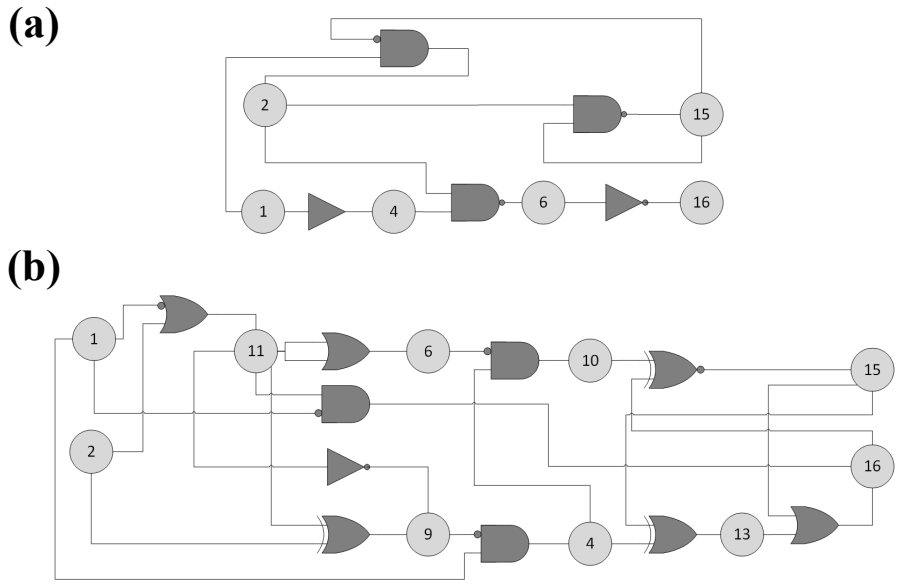

Figure 2.5: Markov brains evolve alternative circuits to encode a motion detection circuit (duplicated logic gates with same inputs and outputs are omitted). (A) Example simple evolved motion detection circuit. (B) Example complex evolved motion detection circuit. Gate symbols are US Standard.

To gain a better understanding of the diversity of neuronal circuits evolved in this study, we performed gate-knockout assays on all 75 brains. We sequentially eliminated each logic gate and

re-measured the mutant brain's fitness, thus allowing us to estimate which gates were essential to the motion detection function (if mutant fitness decreased) and which gates were redundant to the motion detection function (if mutant fitness was equal to the ancestral fitness). There was a wide distribution in the number of essential logic gates, ranging from two logic gates to ten logic gates, with a mean of 4.82 gates (Fig. 2.6(A)). This result supports the idea that there is a wide diversity of possible motion detection circuits available to evolution. We also measured the number of redundant logic gates and found our evolved brains possessed an even greater number of gates that had no apparent contribution to the circuit's function (Fig. 2.6(B)), suggesting that either a large portion of the complexity of these motion detection circuits evolved neutrally, or that selection for redundancy and mutational robustness is involved.

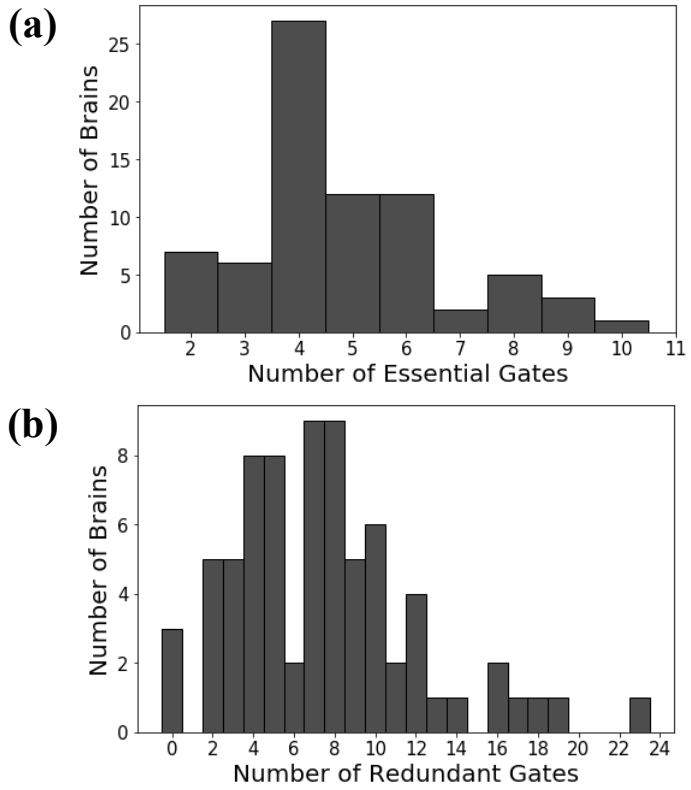

Figure 2.6: Evolved motion detection circuits vary greatly in complexity. (A) Histogram of the number of essential gates (i.e., gates that resulted in a fitness loss when removed) for each evolved motion detection circuit. (B) Histogram of the number of redundant gates (i.e., gates that resulted in no fitness loss when removed) for each evolved circuit.

We also examined the types of logic gates that were either essential or redundant to each brain by

recording the average number each gate was found within each evolved brain. We found surprising similarities in the distribution of the average presence of each logic gate between both essential gates (Fig. 2.7(A)) and redundant gates (Fig. 2.7(B)). The six most-abundant logic gates in both the essential and the redundant gate distribution were NOR, OR-NOT, AND-NOT, NOT, COPY, and EQU. These results suggest either that evolved motion detection circuits may incorporate whichever gates are most easily-evolved (in the sense that they interact with other gates without fitness trade-offs) or that they may have evolved the same redundant gates as their essential gates in order to encode robustness against mutations.

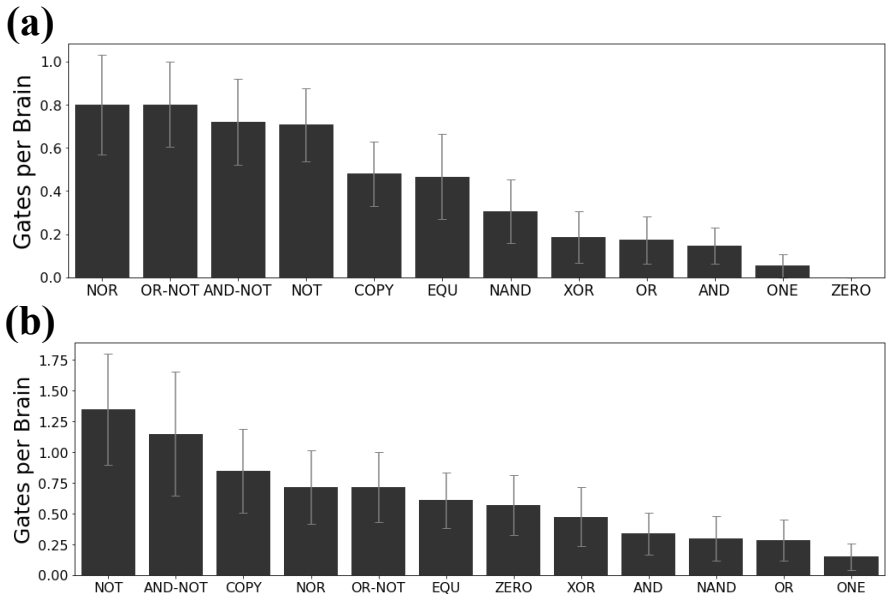

Figure 2.7: Distribution of specific gates used in evolved motion detectors. (A) Average number of essential logic gates of each type of logic gate per evolved brain. Error bars represent 95% confidence intervals. (B) Average number of redundant logic gates of each type of logic gate per evolved brain. Error bars represent 95% confidence intervals.

Multiple pieces of evidence suggest that the complexity of our evolved brains did not evolve solely to perform the motion-detection function. Our evolved brains are more complex than required to encode a motion detection circuit (Fig. 2.6). The large abundance of redundant gates suggests that either these brains are neutrally evolving increased complexity or are evolving mutational robustness due to high mutation rates. The similarities in the distribution of both essential and redundant logic gates suggests either that certain gates arise due to their intrinsic abundance in

the fitness landscape, or because they can compensate for mutations to otherwise essential gates. Therefore, to test for the reason behind our evolved brains' complexity, we hand-wrote a simple Reichardt detector with optimal individual fitness (Fig. 2.8(A)), evolved 100 populations under the same protocol as before, and repeated our knockout analysis. If the evolution of complexity was either non-adaptive or due to selection for increased redundancy and robustness, we would expect these simple brains to increase in complexity upon further evolution. However, if the motion detector circuit's evolved complexity is due to difficult-to-break historical contingency, we would expect little change in the brains evolved from hand-written Reichardt detectors.

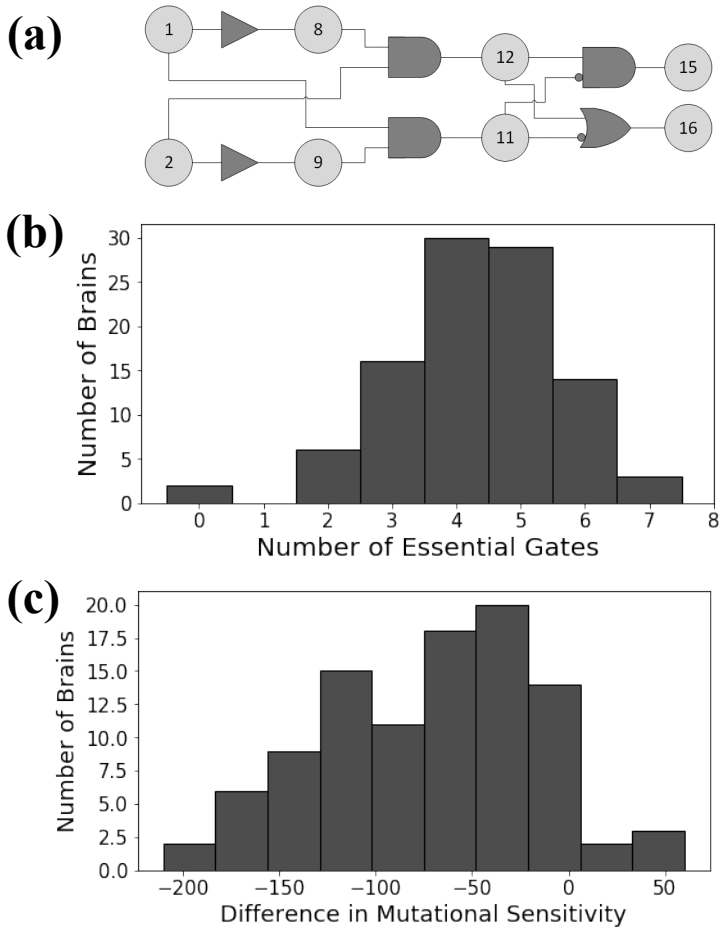

Figure 2.8: Evolution of a simple Reichardt detector leads to greater complexity. (A) Diagram of a hand-written Markov Brain encoding a simple Reichardt detector (B) Distribution of the number of essential gates for brains evolved from a hand-written ancestor. (C) Mutational sensitivity of evolved motion detectors.

The results from the knockout analysis demonstrated that the brains evolved from a hand-written

Reichardt detector increased in complexity when evolved further (Fig. 2.8(B)), suggesting that the increased complexity seen in Fig. 6 was not due to historical contingency, but to other evolutionary factors. To test if these evolved brains were shaped by selection for mutational robustness, we measured the mutational sensitivity of each brain by calculating the average fitness loss from removing one logic gate and multiplying this loss by the total number of gates in each brain. Those evolved brains were less mutationally-sensitive (or more mutationally-robust) than their hand-written ancestor (Fig. 2.8(C)), suggesting that the additional gates evolved in order to increase the brain's robustness to mutations. However, we should also note that some brains did evolve a greater mutational sensitivity, suggesting that either robustness was evolved beyond single-step mutations or that there is some role for non-adaptive evolutionary processes in driving circuit architecture.

2.3 Discussion

We tested if a computational model could evolve a wide diversity of neuronal architectures, and studied evolutionary trends in the evolution of these neuronal architectures. We found that selection for motion detection does lead to a wide diversity of neuronal circuits even though each has the same overall function. Most brains are more complex than the standard model for motion detection: the Reichardt detector. Each brain uses many different logic-gate components, although some gates are more common than others. A large portion of the evolved complexity in these brains results from the evolution of redundant gates. We also showed that even hand-written Reichardt detectors increase in complexity when evolved further, suggesting that the large complexity is due to either non-adaptive evolution or selection for functional redundancy. Measurements of the evolved brains' mutational sensitivity suggested they had indeed evolved mutational robustness, illustrating one additional selective pressure beyond basic functionality on the neuronal architecture of motion-detection circuits.

We undertook this study to see if some of the trends detected in the evolution of genetic circuits occurred in the evolution of Markov brains [194]. As found in many other functional systems, including those based on biochemistry [200] and those based on various digital substrates [157, 50,

30], there is a wide variety of diverse neuronal architectures that can encode a motion-detection circuit that is logically equivalent to that of a Reichardt detector. Our results are in accordance with previous results that showed neuronal circuits with the same functional output could vary between species [171]. These results suggest that a diversity of neuronal architectures may exist for species across life. Our results also suggest that any system with interacting individual components that, when combined, lead to a functioning circuit may possess a diversity of circuits that provide the same function.

While it is perhaps not surprising that our evolved digital brains are different from the default Reichardt detector encoding, we did not expect them to be much more complex. Thus, it is worth discussing how some of our experimental design decisions could have influenced these differences. One likely difference between our evolved brains and real brains is the lack of any fitness cost for larger brains in our model. If each neuron or logic gate was associated with a fitness cost, then one would intuitively expect the evolved brains to be simpler than what we found them to be. On the other hand, neuro-anatomical evidence has suggested that wiring length and connection cost do not appear to be minimized in brains [see also [68]].

Another difference between digital and biological brains is that we only selected on one trait here. The evolution of neuronal circuits is likely constrained by pleiotropic interactions with other functional circuits, as with genetic systems [174]. Finally, compared to biological systems, Markov brains evolved under of a very high mutation rate, something that is known to alter the evolution of genetic architecture towards mutational robustness [203]. It is likely that Markov brains would have evolved less-complex circuits with a decreased mutation rate, although the magnitude of this effect is not known.

We envision the results we presented here as a first step in establishing Markov brains as a model system to study the potential neuronal architectures evolved by Darwinian natural selection. Some of the limitations discussed above present fruitful avenues for future work that may lead to further insights into the evolutionary potential of biological brains. Although we did not attempt a more-precise classification of our evolved circuits beyond their complexity and their specific logic

gates, we see this as a possible endeavor. If the addition of further selection pressures results in the evolution of simpler brains than those evolved here, this task should be achievable. Such studies should lead to a more predictable theory of the diversity of neuronal circuits.

CHAPTER 3

FLIES AS SHIP CAPTAINS? DIGITAL EVOLUTION UNRAVELS SELECTIVE PRESSURES TO AVOID COLLISION IN DROSOPHILA

3.1 Introduction

How animals make decisions has always been an interesting, yet controversial, question to scientists [125] and philosophers alike. Animals obtain various types of sensory information from the environment and then process these information streams so as to take actions that benefit them in survival and reproduction. The visual system plays an important role in providing animals information about their environment, for example when foraging for food, detecting predators or prey, and when searching for potential mates. One of the primary components of visual information is motion detection. Motion is a fundamental perceptual dimension of visual systems [20] and is a key component in decision making in most animals. Here, we study a very particular type of motion detection and concomitant behavior (collision avoidance) in *Drosophila melanogaster* (the common fruit fly), and attempt to unravel the selective (i.e., evolutionary) pressures that might have given rise to this behavior.

D. melanogaster shows a striking difference in behavior when exposed to two different types of optical flow. [25] recorded the interaction of groups of fruit flies in a planar covered arena (so that they could only walk, not fly) and used computer vision algorithms to analyze the walking trajectories in order to study fly behavior. Their analysis revealed that female fruit flies stop walking when they perceive another fly's motion from back-to-front in their visual field (an optical flow referred to as "regressive motion") whereas they keep walking when perceiving conspecifics moving from front-to-back in their visual field (referred to as "progressive motion," see Figure 3.1). [207] further investigated this behavior and tested the "regressive motion saliency" hypothesis, suggesting that flies stop walking when perceiving regressive motion. They used a programmable fly-sized robot interacting with a real fly to exclude other sensory cues such as image expansion

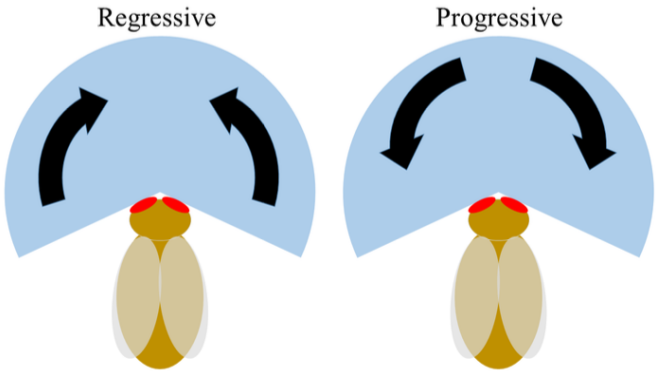

Figure 3.1: An illustration of regressive (back-to-front, left) and progressive (front-to-back, right) optic flows in a fly's retina.

("looming," see [163]) and pheromones. Their results provide rigorous support for the regressive motion saliency hypothesis.

Subsequently, [33] coined the term "generalized regressive motion" for optic flows in which images move clockwise on the left eye and conversely, counterclockwise on the right eye (see Figure 3.1). They presented a geometric analysis for two flies moving on straight, intersecting trajectories with constant velocities and showed that the fly that reaches the intersection first always perceives progressive motion on its retina, whereas the one that reaches the intersection later perceives regressive motion at all times before the other fly reaches the intersection. They went on to suggest that this behavior is a strategy to avoid collisions during locomotion similar to the rules that ship captains use when moving on intersecting paths (see, e.g., [110]).

As intriguing as this hypothesis may seem, it is not clear *a priori* which selective pressures or environmental circumstances could give rise to this behavior. For example, it is unclear whether collision avoidance provides a significant enough fitness benefit. As a consequence, it is possible that the behavior has its origin in a completely different cognitive constraint that is fundamentally unrelated to collision avoidance, or to the rules that ship captains use to navigate the seas. While such questions are difficult to answer using traditional behavioral biology methods, Artificial Life offers unique opportunities to test these hypotheses directly.

In this study, we tested whether collision avoidance can be a sufficient selective pressure for

the described behavior to evolve. We also investigated the environmental conditions under which this behavior could have evolved, in terms of the varying costs and benefits involved. By using an agent-based computational model (described in more detail below), we studied how the interplay (and trade-offs) between the necessity to move and the avoidance of collisions can result in the evolution of regressive motion saliency in digital flies.

Digital evolution is currently the only technique that can study hypotheses concerning the selective pressures necessary (or even sufficient) for the emergence of animal behaviors, as experimental evolution with animal lines of thousands of generations is impractical. In digital evolution, we can study the interplay between multiple factors such as selective pressures, environmental conditions, population size and structure, etc. For example, Olson et al. ([140]) used digital evolution to show that predator confusion is a sufficient condition to evolve swarming behavior, but they also found that collective vigilance can give rise to gregarious foraging behavior in group-living organisms [137]. In principle, any one hypothesis favoring the emergence of behavior can be tested in isolation, or in conjunction [137].

3.2 Methods

3.2.1 Markov Networks

We use an agent-based model to simulate the interaction of walking flies with moving objects (here, potentially conspecifics) in a two-dimensional virtual world. Agents have sensors to perceive their surrounding world (details below) and have actuators that enable them to move in the environment. Agent brains in our experiment have altogether twelve sensors, three internal processing nodes, and one output node (the actuator). The brain controlling the agent is a "Markov network brain" (MNB), which is a probabilistic controller that makes decisions based on sensory inputs and internal nodes [45]. Each node in the network (i.e., sensors, internal nodes, and actuators) can be thought of as a digital (binary) neuron that either fires (value=1), or is quiescent (value=0). Nodes of the network are connected via Hidden Markov Gates (HMGs) that function as probabilistic logic gates. Each HMG is specified by its inputs, outputs, and a state transition table that specifies the

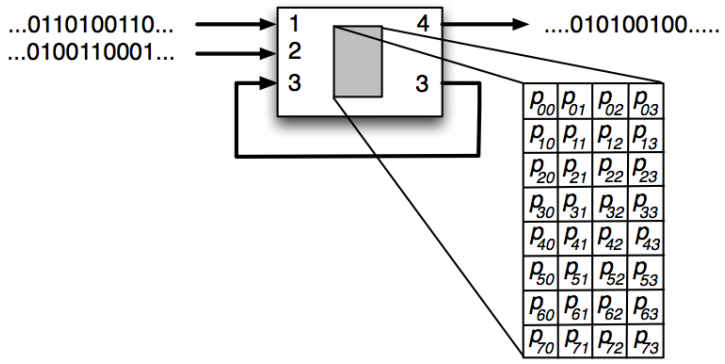

Figure 3.2: Probabilistic logic gates in Markov network brains with three inputs and two outputs. One of the outputs writes into one of the inputs of this gate, so its output is "hidden." Because after firing all Markov neurons automatically return to the quiescent state, values can only be kept in memory by actively maintaining them. Probability table shows the probability of each output given input values.

probability of each output state based on input states (Figure 3.2). For example, in the transition table of Figure 3.2 (a three-input, two-output gate), the probability p_{73} controls the likelihood that the output state is 3 (the decimal equivalent of the binary pattern 11, that is, both output neurons fire) given that the input happened to be state 7 (the decimal translation of 111, i.e., all inputs are active). MNBs can consist of any number of HMGs with any possible connection arrangement, given certain constrains (see for example [45]).

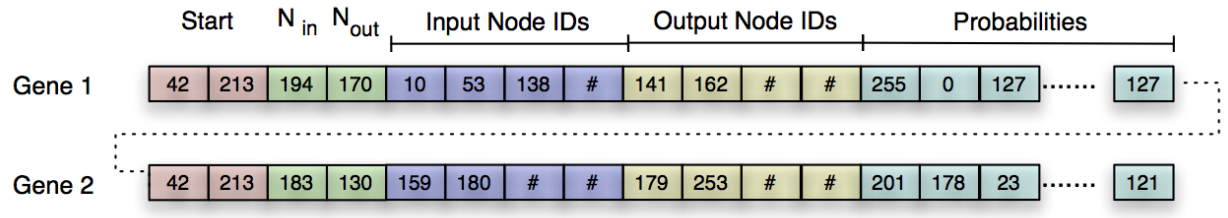

Figure 3.3: An illustration of a portion of genome containing two genes that encode two HMGs. The first two loci represent start codon (red blocks), followed by two loci that determine the number of inputs and outputs respectively (green blocks). The next four loci specify which nodes are inputs of this gate (blue blocks) and the following four specify output nodes (yellow blocks). The remaining loci encode the probabilities of HMG's logic table (cyan blocks).

The number of gates, their connections, and how they work is subject to evolution and changes across individuals and through generations. For this purpose, the agent's brains are encoded in a

genome, which is an ordered sequence of integers, each in the range [0,255], i.e., one byte. Each integer (or byte) is a locus in the genome and specific sequences of loci construct genes, where each gene codes for one HMG. The "start codon" for a gene (i.e., the sequence that determines the beginning of the gene) in our encoding is the pair (42,213) (these numbers are arbitrary). Each gene encodes exactly one HMG, for example as shown in Figure 3.3. The gene specifies the number of inputs/outputs in each HMG, which nodes it reads from and writes to (the connectivity) and the probability table that determines the gates' function. As shown in Figure 3.3, the first two bytes are the start codon, followed by one byte that specifies the number of inputs and one byte for the number of outputs. The bytes are modulated so as to encode the number of inputs and outputs unambiguously. For example, the bytes encoding the number of inputs is an integer in [0,255] whereas a HMG can take a maximum of four inputs, thus we use a mapping function that generates a number $\in$ [1,4] from the value of this byte. The next four bytes specify the inputs of the HMG, followed by another four bytes specifying where it writes to. The remaining bytes of the gene are mapped to construct the probabilistic logic gate table. MNBs have been used extensively in the last five years to study the evolution of navigation [45, 84], the evolution of active categorical perception [116, 9], the evolution of swarming behavior as noted earlier, as well as how visual cortices [34] and hierarchical groups [71] form. In this work, we force the gates to be deterministic rather than probabilistic (all values in the logic table are 0 or 1), which turns our HMGs into classical logic gates.

3.2.2 Experimental Configurations

We construct an initial population of 100 agents (digital flies), each with a genome initialized with 5,000 random integers containing four start codons (to jump-start evolution). Agents (and by proxy the genomes that determine them) are scored based on how they perform in their living environment. The population of genomes is updated via a standard Genetic Algorithm (GA) for 50,000 generations, where the next generation of genomes is constructed via roulette wheel selection combined with mutations (detailed GA specifications are listed in Table 3.1). To control

for the effects of reproduction and similar effects, there is no crossover or immigration in our GA implementation.

Each digital fly is put in a virtual world for 25,000 time steps, during which time its fitness score is evaluated. During each time step in the simulation, the agent perceives its surrounding environment, processes the information with its MNB, and makes movement decisions according to the MNB outputs. The sensory system of a digital fly is designed such that it can see surrounding objects within a limited distance of 250 units, in a 280° pixelated retina shown in Figure 3.4. The state of each sensor node is 0 (inactive) when it does not sense anything within the radius, and turns to 1 (active) if an object is projected at that position in the retina. Agents in this experiment have one actuator node that enables them to move ahead or stop, for active (firing) and non-active (quiescent) states respectively.

Table 3.1: Configurations for GA and Environmental setup

GA Parameters		**Environment Parameters**	
Population size	100	Vision range	250
Generations	50,000	Field of vision	280°
Point mutation rate	0.5%	Collision range	60
Gene deletion rate	2%	Agent velocity	15
Gene duplication rate	5%	Event time steps	250
Initial genome length	5,000	No. of events	100
Initial start codons	4	Moving reward	0.004
Crossover	None	Collis. penalty	1,2,3,5,10
Immigration	None	Replicates	20

In our experiment, the digital flies exist in an environment where they should move to gain fitness, representing the fact that organisms should forage for resources, mates, and avoiding predators. Thus, the fitness function is set so that agents are rewarded for moving ahead at each update of the world, and are penalized for colliding with objects. The amount of fitness they gain for moving (the benefit) is characteristic of the environment, and we change it in different treatments. The penalty for collisions represents the importance of collision avoidance for their survival and reproduction, and we vary this cost also. Each digital fly sees 100 moving objects (one at a time) during its lifetime, and we say that it experiences 100 "events." The penalty-reward ratio (PR) determines

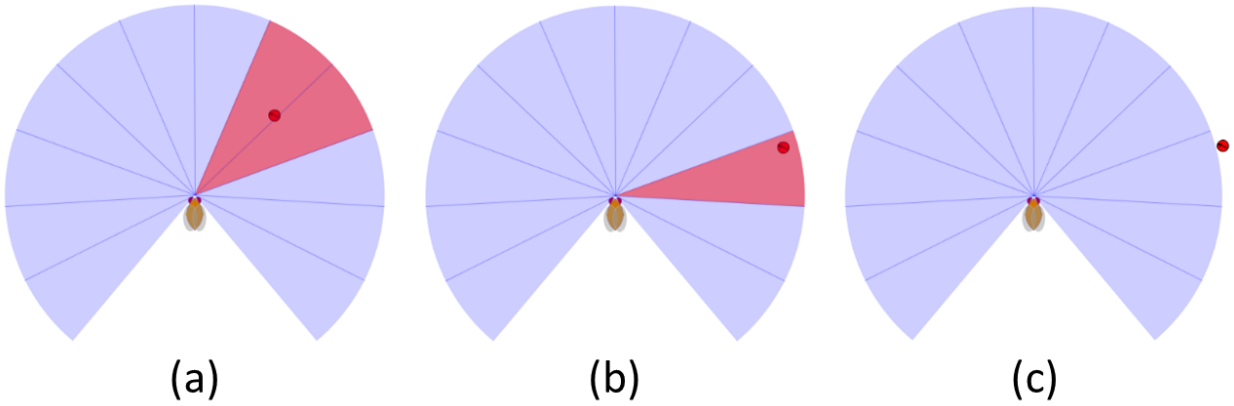

Figure 3.4: The digital fly and its visual field in the model. Flies have a 12 pixel retina that is able to sense surrounding objects in 280° within a limited distance (250 units). The red circle is an external object that can be detected by the agent within its vision field. Activated sensors are shown in red, while inactive sensors are blue. In (A) the object activates two sensors, in (B) the object is detected in one sensor, and in (C) the object is outside the range.

the amount of penalty of collision divided by the reward for moving during the entirety of an event. So for example, PR=1 means the agent loses all the rewards it gained by walking during the whole event if it collides with the object in that event:

$$\text{fitness} = \sum_{\text{events}} (\text{reward} - PR \times \text{collision}) \ , \tag{3.1}$$

where reward $\in [0, 1]$ reflects how many time steps the agent moved during the event. Our experiments are constructed such that *all* objects that produce regressive motion in the digital retina *will* collide with the fly if it keeps moving. The reason for biasing our experiments in this manner is explained in the following section.

3.2.3 Collision Probability in Events with Regressive Optic Flow

As mentioned earlier, Chalupka et al. ([33]) showed that for two flies moving on straight, intersecting trajectories with constant velocities, the fly that reaches the intersection first always perceives progressive motion on its retina while the counterpart that reaches the intersection later perceives regressive motion at all times before the first fly reaches the intersection. However, this does *not* imply that all objects that produce a regressive motion on a fly's retina will necessarily collide with

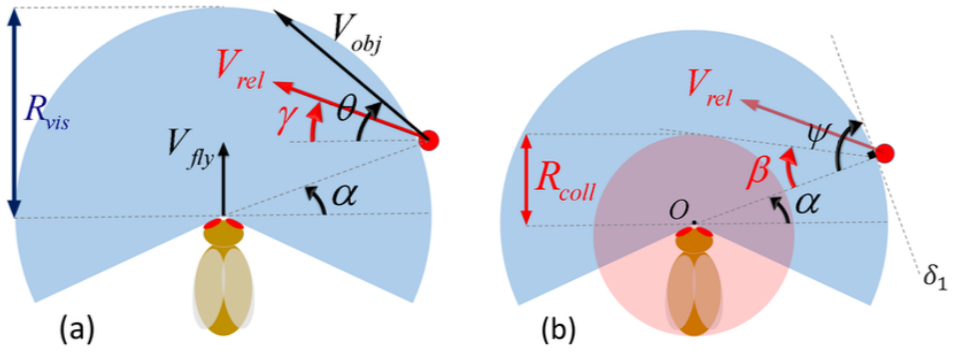

Figure 3.5: An illustration of a moving fly at the onset of the event.

it. In this section we present a mathematical analysis to discover how often objects that produce regressive motion in the fly's retina will eventually collide with the fly if it continues walking.

Suppose a fly moves on a straight line with constant velocity $\boldsymbol{V_{fly}}$ and an object is also moving on a straight line with constant velocity $\boldsymbol{V_{obj}}$ (Figure 3.5(A)). The fly is able to perceive objects within distance R_{vis}, its vision range (Figure 3.5(A)). The object is assumed to be a point in the plane and the distance between this point and the center of the visual field of the fly is defined to be the distance between them. We define "the onset of the event" as the first time the object is detected by the fly. At the onset of the event, the object is at the distance R_{vis} of the fly at relative azimuthal angle $\alpha \in [0, \frac{\pi}{2}]$ (Figure 3.5(A)). We assume that the object can be at any relative position $\boldsymbol{R_{vis}} = (R_{\text{vis}}, \alpha)$[1] with equal probabilities (the probability distribution of α is uniform around the fly). The velocity of the object can be represented as $\boldsymbol{V_{obj}} = (V_{\text{obj}}, \theta)$ where $\theta \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ (note that $\boldsymbol{V_{obj}}$ is constant). We also assume that the velocity of the object can point in all directions with equal probabilities (the probability distribution of θ is uniform). The relative velocity of the object with respect to the fly is $\boldsymbol{V_{rel}} = \boldsymbol{V_{obj}} - \boldsymbol{V_{fly}}$ (Figure 3.5). Since both $\boldsymbol{V_{obj}}$ and $\boldsymbol{V_{fly}}$ are constant, $\boldsymbol{V_{rel}}$ is also a constant vector.

[1]Here and below, we represent vectors either in boldface or by the parameters that determine them within a planar polar coordinate system. Thus the vector $\boldsymbol{R}$ is represented by $(|\boldsymbol{R}|, \phi)$, where $R_x = R\cos\phi$ and $R_y = R\sin\phi$.

3.2.3.1 Proposition 1.

A moving object produces regressive motion on a fly's retina if:

$$\theta > -\alpha + \arcsin(\frac{V_{\text{fly}}}{V_{\text{obj}}} \cos\alpha) \ . \tag{3.2}$$

3.2.3.2 Proof.

In order for the object to produce regressive motion on the retina, the relative velocity should be pointed above the center point O. The relative velocity direction γ can be found awith $\boldsymbol{V_{\text{rel}}} = (V_{\text{rel}}, \gamma)$, as

$$\gamma = \arctan\left(\frac{V_{\text{rel}_y}}{V_{\text{rel}_x}}\right) = \arctan\left(\frac{V_{\text{obj}} \sin\theta - V_{\text{fly}}}{V_{\text{obj}} \cos\theta}\right) \ . \tag{3.3}$$

The angle γ should be greater than the central angle (Figure 3.5(B)), that is, $\gamma > -\alpha$. Replacing γ and simplifying, we obtain:

$$\theta > -\alpha + \arcsin(\nu \cos\alpha), \quad \nu = \frac{V_{\text{fly}}}{V_{\text{obj}}} \ . \tag{3.4}$$

For smaller values of θ, the object produces progressive optic flow. We thus define $\theta_{\text{min}} = -\alpha + \arcsin(\nu \cos(\alpha))$ as the minimum angle θ that produces regressive motion on the retina.

3.2.3.3 Definition 1.

The object remains "observable" to the fly after the onset of the event if its relative velocity is directed toward the inside of the fly's vision field (to the left of the tangent line δ_1 in Figure 3.5(B)).

3.2.3.4 Proposition 2.

The object remains observable to the fly if:

$$\theta < \arccos(-\frac{V_{\text{fly}}}{V_{\text{obj}}} \sin\alpha) - \alpha \ . \tag{3.5}$$

3.2.3.5 Proof.

According to the definition the sufficient condition for observability is that γ should be less than the tangent line δ_1 angle: $\gamma < -\alpha + \frac{\pi}{2}$. Replacing γ and simplifying we obtain

$$\theta < \arccos(-\nu \sin \alpha) - \alpha \ . \tag{3.6}$$

For greater values of θ, the object will be out of vision range of the fly. Thus the maximum value that θ can take on is:

$$\theta_{\max} = \arccos(-\nu \sin \alpha) - \alpha \ . \tag{3.7}$$

In order for the object to produce regressive motion on fly's retina and also remain observable to the fly, relative velocity should be within the arc ψ (Figure 3.5(B)).

3.2.3.6 Definition 2.

The object collides with the fly if its distance with the fly is less than "collision range" R_{coll} (Figure 3.5(B)).

3.2.3.7 Proposition 3.

An object that creates regressive optic flow on the fly's retina and remains observable will collide with it if:

$$\theta < \phi + \arcsin(\nu \cos \phi), \quad \phi = \arcsin(\frac{R_{\text{coll}}}{R_{\text{vis}}}) - \alpha \ . \tag{3.8}$$

3.2.3.8 Proof.

The relative velocity of such object is within arc ψ. This object will collide with the fly if its relative velocity is within the arc spanned by the angle β, i.e. lower than tangent line to collision circle (Figure 3.5(B)). This condition holds true if:

$$\gamma < \beta - \alpha, \quad \beta = \arcsin(\frac{R_{\text{coll}}}{R_{\text{vis}}}) \ . \tag{3.9}$$

Let $\rho = \frac{R_{\text{coll}}}{R_{\text{vis}}}$ and $\phi = \beta - \alpha$. Replacing γ and rearranging gives:

$$\theta < \phi + \arcsin(\nu \cos \phi) \ . \tag{3.10}$$

For greater values of θ, the object produces regressive motion on the fly's retina but does not collide with it. So the threshold collision angle is given by:

$$\theta_{\text{col}} = \phi + \arcsin(\nu \cos \phi) \ . \tag{3.11}$$

As mentioned, we assume that the probability distribution of the direction of the object velocity, θ is uniform.

3.2.3.9 Definition 3.

For an object at initial position α, the probability Π_{coll} is the range of velocity directions θ such that the object collides with the fly divided by the range of directions with which it creates regressive optic flow on fly's retina (see Figure 3.5(B)):

$$\Pi_{\text{coll}}(\alpha, \nu, \rho) = \frac{\theta_{\text{col}} - \theta_{\text{min}}}{\theta_{\text{max}} - \theta_{\text{min}}} \ . \tag{3.12}$$

Integrating this function over the range of possible initial relative positions, the probability that an event results in a collision given that the object produces regressive motion on an fly's retina can be found as:

$$\Pi_{\text{coll}}(\nu, \rho) = \int_{\alpha_{\text{min}}}^{\alpha_{\text{max}}} \Pi_{\text{coll}}(\alpha, \nu, \rho) d\alpha \ , \tag{3.13}$$

where α_{min} is either 0 or the minimum value of α for which there exists a θ with which the object can produce a regressive motion on fly's retina, and α_{min} is either 90 or maximum value of α for which there exists a θ with which the object remains observable to the fly.

We calculated the integral (3.13) numerically and show the results in Figure 3.6 for different values of fly-object velocity ratios ν and different collision range-vision range ratios ρ. As can be seen from Figure 3.6, for R_{vis}=60 mm [207] and R_{coll}=15 mm (our assumption), the collision

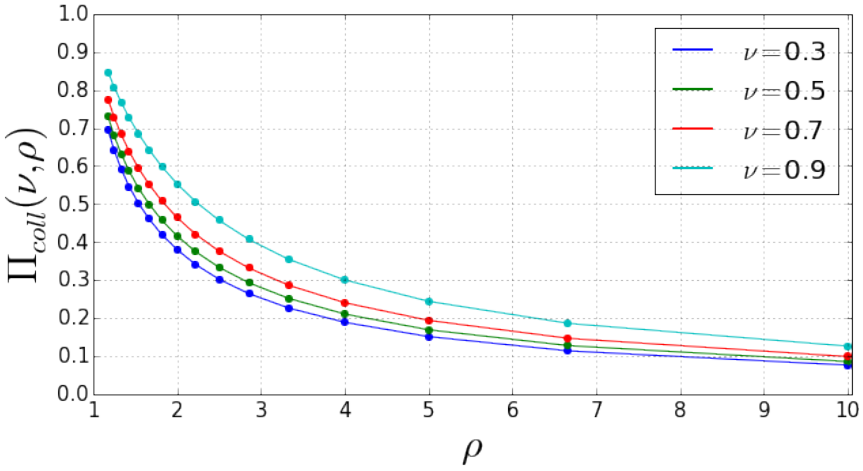

Figure 3.6: Probability of collision $\Pi_{\text{coll}}(\nu, \rho)$ with an object that creates regressive motion on the retina as a function of the ratio of vision radius to collision radius ρ, for different fly-object velocity ratios ν.

probability is around 0.2-0.3. This implies that if encounters are created randomly, regressive motion on the retina is not predictive of collision, and as a consequence it is unreasonable to expect that digital evolution will produce collision avoidance in response, as only 1 in 5 to 1 in 3 regressive motions actually lead to collisions. This was borne out in experiments, and we thus decided to bias the events in such a manner that *all* events that leave a regressive motion signature in the retina will lead to collision. Note that this is not necessarily an unrealistic assumption, as we have not analyzed a distribution of realistic "events" (such as is available in the data set of [25]). It could very well be that the way real flies approach each other differs from the uniform distributions that went into the mathematical analysis presented here.

3.3 Results

We conducted experiments with five different fitness functions representing different environments. Environments differ in the amount of fitness individuals gain when moving and in the penalty incurred by a collision. Evolved agents use various strategies to avoid collisions and maximize the travelled distance, but one of the most successful strategies they use is indeed to categorize visual cues into regressive and progressive optic flows. We find that agents categorize these visual cues only in some regions of the retina: the regions in which collisions take place more frequently.

They then use this information to cast a movement decision: they keep moving when seeing an object creating progressive optic flow on their retina, and stop when the object creates regressive optic flow on their retina. However, they do not stop for the entire duration of the event, i.e., the whole time they perceive regressive optic flow. Rather they stop during only a portion of the event, which helps the agent to avoid a collision with the object while maximizing their walking duration and hence gaining higher fitness.

The strategy of using regressive motion as a cue for collision [33], similar to the observed behavior in fruit flies [207] evolves in our experimental setup under some environmental circumstances (discussed below). We refer to this strategy as regressive-collision-cue (RCC) and we define it in our experimental setup as follows:

1) The moving object produces *regressive* motion on the agent's retina during an event and the agent stops at least for some time during that event, or

2) The moving object produces *progressive* motion on the agent's retina during an event and the agent does not stop during that event. The number of events (out of 100) in which the agent uses this strategy is termed the "RCC value."

We now discuss the results of an experiment in which the RCC strategy has evolved. We take the most successful agent at the end of that experiment and analyze its behavior. This agent evolved in an environment with penalty-reward ratio of 2, meaning the penalty of each collision equals twice the maximum reward the agent can gain in 2 events. Figure 3.7 shows whether the agent stopped during an event, stop probability (blue triangles), as a function of the angular velocity of the image on the agent's retina for 100 events. In that figure, the angular velocity of the image on agent's retina is negative for regressive optic flow and positive for progressive events. Simulation units are converted to plotted values (in deg/s and mm/s) by equalizing dimensionless values ν, and ρ in simulation and actual values: R_{vis}=60 mm [207], V_{fly}=20 mm/s [207], R_{coll}=15 mm (our assumption). We can see from the figure that out of all 100 events, the agent did not stop during one event with regressive motion while for two progressive events, it stopped. In the remaining events the agent accurately uses the RCC strategy (resulting in an RCC value=97). The average velocity of

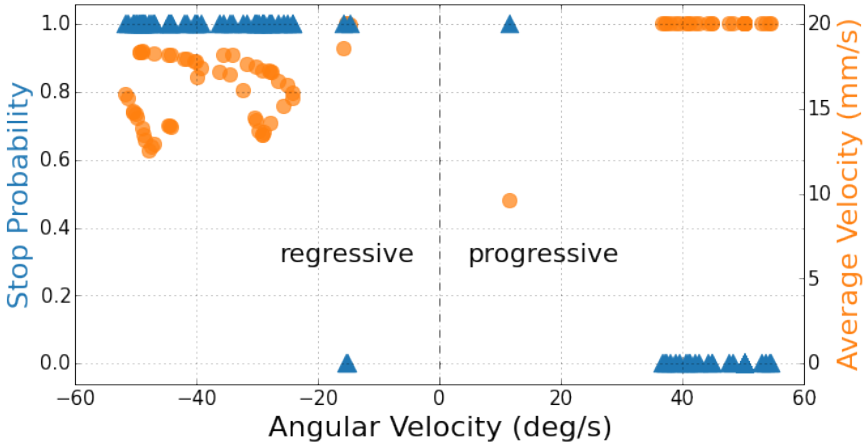

Figure 3.7: The stop probability of the evolved agent vs. the angular velocity of the image on its retina for 100 events. Positive values of angular velocity show progressive motion events and negative angular velocities stand for regressive motion events. The average velocity of the agent is also shown during each event.

the agent during each event is also shown (solid orange circles), which reflects the number of time steps the agent moves during that event (and thus indirectly how often it stops). For progressive motions, the stop probability is zero (the agent continues to move during the event) and thus the velocity of the agent is maximal during that event. For regressive optic flow (negative angular velocities), the average velocity during each event is less than maximum and for extreme angular velocities, as it only needs to stop for shorter durations to avoid collisions.

In order to quantitatively analyze how using regressive motion as a collision cue benefits agents to gain more fitness, we traced this particular agent's evolutionary line of descent (LOD) by following its lineage backwards for 50,000 generations mutation by mutation until we reached the random agent that we used to seed the initial population (see [104] for more details on how to construct evolutionary lines of descent for digital organisms). Figure 3.8 shows the fitness and the RCC value vs. generation for this agent's LOD. It is evident from these results that evolving this strategy benefits agents in gaining fitness compared to the rest of the population in this environment as high peaks of fitness occur at high RCC values and conversely, the fitness drops as the RCC value decreases. Nevertheless, this strategy does not evolve all the time. Figure 3.9 shows the fitness and RCC for all 20 replicates in the environment with penalty-reward ratio of 2. We can see that the

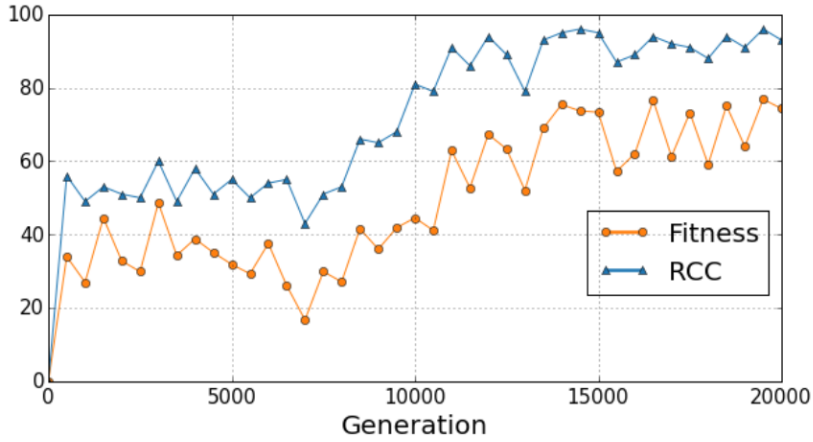

Figure 3.8: Fitness and regressive-collision-cue (RCC) value on the line of descent for an agent that evolved RCC as a strategy to avoid collisions. Only the first 20,000 generations are shown, for every 500 generations.

mean fitness of all 20 replicates is around 20% less than the fitness of the agent that evolved the RCC strategy. The mean RCC value for all 20 replicates is also $\approx$ 20% less than that of an agent that evolved the RCC strategy.

The difficulty to evolve the RCC strategy is not limited to the number of runs in which this behavior evolved out of all replicates in some environment (we also tried running the experiment for longer evolutionary times but the results do not change significantly). Environmental conditions also play a key role in the evolution of this behavior. Figure 3.10 shows the RCC value distribution for 20 replicates in five different environments. In order to calculate the RCC value in each replicate, we took the average of the RCC value in the last 1,000 generations on the line of descent to compensate for fluctuations. We observe that the RCC strategy only evolves in a narrow range of penalty-reward ratio, namely for PR=2 and PR=3. According to Figure 3.10, higher values of penalty on the one hand discourage the agents from walking in the environment (they simply choose to remain stationary), and therefore prevent them from exploring the fitness landscape. Lower values for the penalty, on the other hand, result in indifference to collisions and thus, the optimal strategy (probably the local optimum) in these environments is to keep walking and ignore all collisions. For lower values of the penalty, the RCC value is $\approx$ 55%, which means they evolve

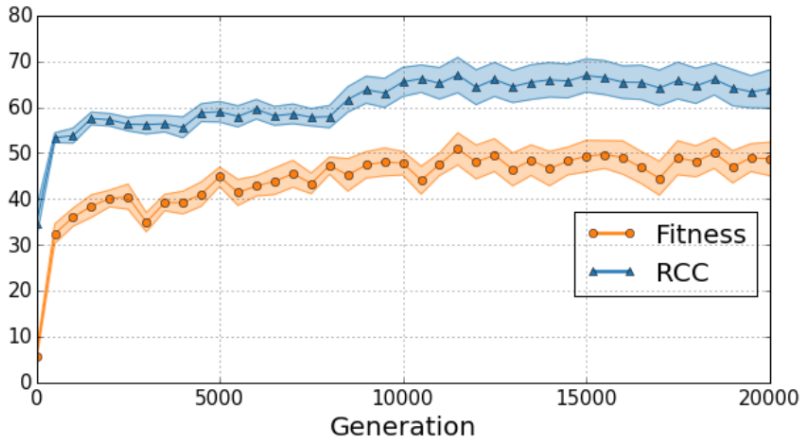

Figure 3.9: Mean values of fitness and regressive-collision-cue (RCC) over all 20 replicates vs. evolutionary time in the line of descent in the environment with penalty-reward ratio of 2. Standard error lines are shown with shaded areas around mean values. Only the first 20,000 generations are shown, for every 500 generations.

to stop in obvious cases that end up in collision (if they keep moving, the RCC value should be 50).

3.4 Discussion

We used an agent-based model of flies equipped with MNBs that evolve via a GA to study the selective pressures and environmental conditions that can lead to the evolution of collision avoidance strategies based on visual information. We specifically tested cognitive models that invoke "regressive motion saliency" and "regressive motion as a cue for collision" to understand how flies avoid colliding with each other in two-dimensional walks. We showed that it is possible to configure the experiment in such a manner that "regressive-collision-cue" (RCC) evolves as a strategy to avoid collisions. However, the conditions under which the RCC strategy evolved in our experiments are limited: the strategy only evolved in a narrow range of environmental conditions and even in those environments, it does not evolve all the time. In addition, we showed that from general principles, only a small percentage of events in which an agent perceives regressive optical flow eventually leads to a collision, so that RCC as a sole strategy is expected to have a large false positive rate, leading to unnecessary stops.

As discussed in the Methods section, our experimental implementation is biased in such a way

that *all* regressive motion events lead to a collision if the agent does not stop during that event. If the moving object's velocity direction is distributed uniformly randomly in all directions, the probability that a regressive event ends up in a collision is rather low ($\approx 20\%$ in our implementations). Because the false positive rate of using regressive optical flow as the only predictor of collisions is liable to thwart the evolution of an RCC strategy, we biased our setup in such a way that the false-positive rate is zero, a bias that does not significantly influence the outcome of our experiments. Consider an environment in which only a percentage of events with regressive motion end up in collision. This is similar to an environment with a lower penalty for collisions (as long as the strategy evolves at all) since the agent's fitness is scored at the end of its lifetime (all 100 events) not during each event.

However, there is a difference between a lower percentage of collisions in regressive events and lower penalty for collisions, namely a lower probability of collision in regressive motion events is equivalent to a higher amount of noise in the cue that the agent takes from the environment, compared to the case of lower penalties for collision. In other words, if 100% of all regressive motion events lead to collisions, the agent associates regressive motion events with collisions with certainty. Thus, implementing the experiments with 100% collisions in regressive motion events is tantamount to eliminating the noise in sensory information, which generally aids evolution. Compensating for noise in sensory information could also be achieved if we scored agents in every single event, and informed them about their performance in that event (feedback learning). We did not use feedback learning here, but plan to do so in future experiments.

We conclude that the evolution of "regressive motion saliency" is unlikely to have happened only due to collision avoidance as the selective pressure. It is important to remember that walking is not the most frequent activity in fruit flies. Further, flies do not usually live in high density colonies and therefore do not find themselves on collision courses very often. It may be the case that components of this strategy (namely categorizing the optic flow as regressive or progressive) have evolved under different selective pressures entirely unrelated to the present test situation, and was further evolved to enhance collision avoidance with conspecifics while moving (a type

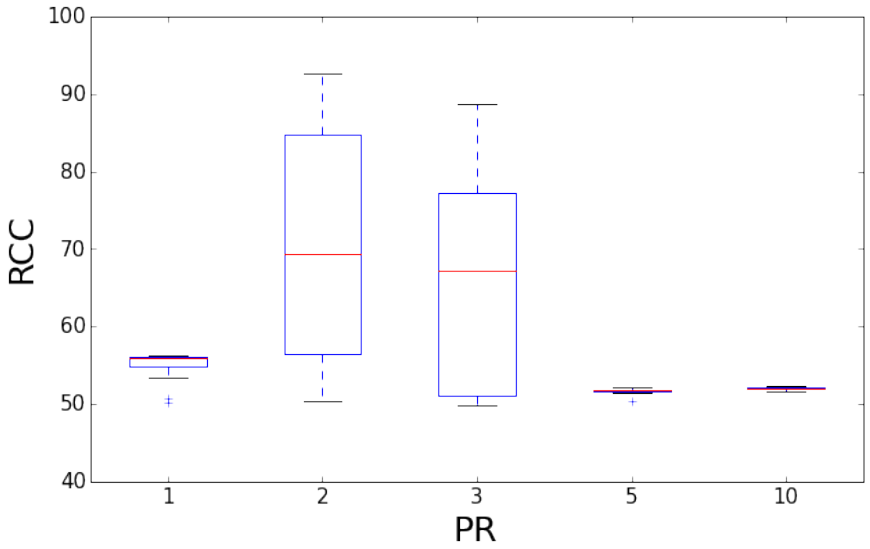

Figure 3.10: RCC value distribution in environments with different penalty-reward ratios. Each box-plot shows the RCC value averaged over the last 1000 generations on the line of descent for 20 replicates.

of exaptation). For example, detecting predators is a strong selective pressure in the evolution of visual motion detection, including the categorization of that cue so as to take appropriate actions. It may be interesting to study the behavior of flies interacting with animals or objects that are not perceived as conspecifics.

CHAPTER 4

CAN TRANSFER ENTROPY INFER INFORMATION FLOW IN NEURONAL CIRCUITS FOR COGNITIVE PROCESSING?

4.1 Introduction

When searching for common foundations of cortical computation, more and more emphasis is being placed on information-theoretic descriptions of cognitive processing [148, 161, 3, 136, 201]. One of the core tasks in the analysis of cognitive processing is to follow the flow of information within the nervous system, by finding cause-effect components. Indeed, understanding causal relationships is considered to be fundamental to all natural sciences [27]. However, inferring causal relationships and separating them from mere correlations is difficult, and the subject of ongoing research [60, 145, 146, 179, 10]. The concept of *Granger causality* is an established statistical measure that aims to determine directed (causal) functional interactions among components or processes of a system. Schreiber [165] described Granger causality in terms of information theory by introducing the concept of *transfer entropy* (TE). The main idea is that if a process X is influencing process Y, then an observer can predict the future state of Y more accurately given the history of both X and Y (written as $X_t^{(k)}$ and $Y_t^{(\ell)}$, where k and ℓ determine how many states from the past of X and Y are taken into account) compared to only knowing the history of Y. According to Schreiber, the transfer entropy $\mathrm{TE}_{X\to Y}$ quantifies the flow of information from process X to Y:

$$\begin{aligned} \mathrm{TE}_{X\to Y} &= I(Y_{t+1} : X_t^{(k)} |\, Y_t^{(\ell)}) = H(Y_{t+1} |\, Y_t^{(\ell)}) - H(Y_{t+1} | Y_t^{(\ell)}, X_t^{(k)}) = \\ &= \sum_{y_{t+1}} \sum_{x_t^{(k)}} \sum_{y_t^{(\ell)}} p(y_{t+1}, x_t^{(k)}, y_t^{(\ell)}) \log \frac{p(y_{t+1} | x_t^{(k)}, y_t^{(\ell)})}{p(y_{t+1} | y_t^{(\ell)})} \,. \end{aligned} \tag{4.1}$$

Here as before, $X_t^{(k)}$ and $Y_t^{(\ell)}$ refer to the history of the processes X and Y, while Y_{t+1} refers to the variable at $t+1$ only. Further, $p(y_{t+1}, x_t^{(k)}, y_t^{(\ell)})$ is the joint probability of Y_{t+1} and the histories $X_t^{(k)}$ and $Y_t^{(\ell)}$, while $p(y_{t+1} | x_t^{(k)}, y_t^{(\ell)})$ and $p(y_{t+1} | y_t^{(\ell)})$ are conditional probabilities.

The transfer entropy (4.1) is a conditional mutual entropy, and quantifies what the process Y

at time $t + 1$ knows about the process X up to time t, given the history of Y up to time t (see [23] for a thorough introduction to the subject). Specifically, $\mathrm{TE}_{X \to Y}$ measures "how much uncertainty about the future course of Y can be reduced by the past of X, given Y's own past." Transfer entropy reduces to Granger causality for so-called "auto-regressive processes" [14] (which encompasses most biological dynamics), and has become one of the most widely used directed information measures, especially in neuroscience (see [199, 202, 201, 23] and references cited therein).

While transfer entropy is sometimes used to infer causal influences between susbsystems, it is important to point out that inferring causal relationships is different from inferring information flow [107]. In complex systems (for example, in computations that a brain performs to choose the correct action given a particular sensory experience) events in the sensory past can causally influence decisions significantly distant in time, and to capture such influences using the transfer entropy concept requires a careful analysis in which not only the history lengths k and ℓ used in Equation (4.1) must be optimized, but false influences due to linear mixing of signals (which can mimic causal influences) must also be corrected for [199, 23]. In some sense, inferring information flow is a much simpler task than finding all causal influences, as we need only to identify (and quantify) the sources of information transferred to a particular variable. More precisely, for this application the pairwise transfer entropy is used to find candidate sources (in the immediate past) that account for the entropy of a particular neuron.

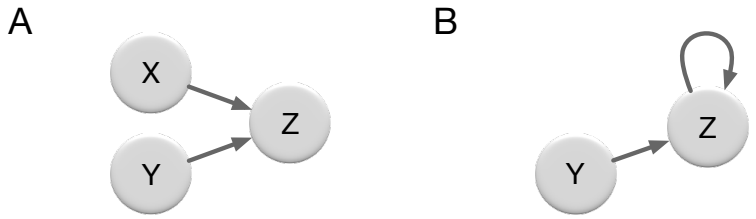

Figure 4.1: (A) A network where processes X and Y influence future state of Z, $Z_{t+1} = f(X_t, Y_t)$. (B) A feedback network in which processes Y and Z influence future state Z, $Z_{t+1} = f(Y_t, Z_t)$.

Using transfer entropy to search for and detect directed information was shown to lead to inaccurate assessments in simple case studies [76, 77]. For instance, James et al. [76] presented two examples in which TE underestimates the flow of information from inputs to output in one

example, and overestimates it in the other. In the first example, they define a simple system with three binary variables X, Y, and Z where $Z_{t+1} = X_t \oplus Y_t$ ($\oplus$ is the exclusive OR logic operation) and variables X and Y take states 0 or 1 with equal probabilities, i.e., $P(X = 0) = P(X = 1) = P(Y = 0) = P(Y = 1) = 0.5$ (this 2-to-1 relation is schematically shown in Figure 4.1A). In this network, $\mathrm{TE}_{X\to Z} = \mathrm{TE}_{X\to Z} = 0$ whereas the entropy of the process Z, $H(Z) = 1$ bit, and variables X and Y certainly influence the future state of Z. In this example, the entropy of Z can be reduced by 1 bit but the TE does not attribute this entropy to either variables X or Y and as a consequence the TE underestimates the flow of information from X and Y to Z. In another example, they define a system with two binary variables Y and Z, where $Z_{t+1} = Y_t \oplus Z_t$ and similar to the previous example, $P(Y = 0) = P(Y = 1) = P(Z = 0) = P(Z = 1) = 0.5$ (this feedback loop relation is schematically shown in Figure 4.1B). In this scenario, $\mathrm{TE}_{Y\to Z} = 1$ bit, which implies that the entire 1 bit of entropy in Z is coming from process Y. However, this is not correct since both Y and Z are equally contributing to determine the future state of Z. In this example, TE overestimates the information flow from process Y to Z. It is also noteworthy that in this example the processed information (defined as $I(Z_t : Z_{t+1})$) vanishes, which again does not correctly detect the other source, Z_t, from which the information is coming. As acknowledged by the authors in [76], expecting that the entropy of the output $H(Z_{t+1})$ is given simply by the sum of the transfer entropy from each of the inputs independently is a naive interpretation of information flow. Indeed, this is generally not the case, even if the two sources are uncorrelated. Consider for example the first system described above in which $Z_{t+1} = f(X_t, Y_t)$. Suppose f is a deterministic function of X_t and Y_t, in which case the conditional entropy $H(Z_{t+1}|X_t, Y_t) = 0$. Then, the entropy $H(Z_{t+1})$ decomposes into the sum of an unconditional and a conditional transfer entropy

$$H(Z_{t+1}) = \mathrm{TE}_{Y\to Z} + \mathrm{TE}_{X\to Z|Y_t} \ , \tag{4.2}$$

where the conditional transfer entropy is defined as (see [23], section 4.2.3)

$$\mathrm{TE}_{Y\to Z|X_t} = I(Y_t : Z_{t+1}|Z_t, X_t) \ . \tag{4.3}$$

Using this definition, it is easy to show that

$$\mathrm{TE}_{Y\to Z} = \mathrm{TE}_{Y\to Z|X_t} + I(X_t : Y_t : Z_{t+1}|Z_t) \ , \tag{4.4}$$

and Equation (4.2) can be rewritten in terms of transfer entropies only, or else conditional transfer entropies only, as

$$H(Z_{t+1}) = \mathrm{TE}_{Y\to Z|X_t} + \mathrm{TE}_{X\to Z|Y_t} + I(X_t : Y_t : Z_{t+1}|Z_t) = \mathrm{TE}_{Y\to Z} + \mathrm{TE}_{X\to Z} - I(X_t : Y_t : Z_{t+1}|Z_t) \tag{4.5}$$

In light of Equation (4.5), it then becomes clear that the naive sum of the transfer entropies $\mathrm{TE}_{X\to Z}$ and $\mathrm{TE}_{Y\to Z}$ (or naive sum of conditional transfer entropies) must fail to account for the entropy of Z whenever the term $I(X_t : Y_t : Z_{t+1}|Z_t)$ is non-zero, and therefore will fail to fully and accurately quantify information transferred from sources X and Y. Therefore, the error in information flow estimate when using transfer entropy is simply given by the absolute value of $I(X_t : Y_t : Z_{t+1}|Z_t)$ (same when using conditional transfer entropies).

Now consider the second example system with a feedback loop in which $Z_{t+1} = f(Y_t, Z_t)$, and again suppose f is a deterministic function which implies $H(Z_{t+1}|Y_t, Z_t) = 0$. In this case, there is a similar information decomposition that now involves a shared entropy $I(Y_t : Z_t : Z_{t+1})$

$$I(Y_t : Z_{t+1}) = \mathrm{TE}_{Y\to Z} + I(Y_t : Z_t : Z_{t+1}) \ . \tag{4.6}$$

Here, the entropy $H(Z_{t+1})$ can be written in terms of transfer entropy and processed information (recall that $H(Z_{t+1}|Z_t, Y_t) = 0$)

$$H(Z_{t+1}) = \mathrm{TE}_{Y\to Z} + I(Z_t : Z_{t+1}) \ . \tag{4.7}$$

While Equation (4.7) shows that the sum of transfer entropy $\mathrm{TE}_{Y\to Z}$ and processed information $I(Z_t : Z_{t+1})$ account for all the entropy Z_{t+1}, these two terms do not always individually identify the sources of information flow correctly. For instance, we have seen that in the second example (where $Z_{t+1} = Y_t \oplus Z_t$) the processed information $I(Z_t : Z_{t+1})$ vanishes even though variable Z_t most definitely influences the state of variable Z_{t+1}. As discussed earlier, all the information transferred

to Z_{t+1} in that case is attributed to variable Y_t. Note that the processed information can be written as

$$I(Z_t : Z_{t+1}) = I(Z_t : Z_{t+1}|Y_t) + I(Z_{t+1} : Z_t : Y_t) \tag{4.8}$$

where $I(Z_t : Z_{t+1}|Y_t) = 1$ and $I(Z_{t+1} : Z_t : Y_t) = -1$.

Note that for the most general case where function f can be non-deterministic and the network with or without feedback loop, the full entropy decomposition can be written as

$$H(Z_{t+1}) = \mathrm{TE}_{Y\to Z|X_t} + \mathrm{TE}_{X\to Z|Y_t} + I(X_t : Y_t : Z_{t+1}|Z_t) + I(Z_{t+1} : Z_t) + H(Z_{t+1}|X_t, Y_t, Z_t) \ . \tag{4.9}$$

There is also another key factor in the examples described above that results in misestimating information flow when using transfer entropy. In both examples, the input to output relation is implemented by an XOR function. For instance, in the first example ($Z_{t+1} = X_t \oplus Y_t$), the transfer entropy $\mathrm{TE}_{X\to Z}$ considers X in isolation and independent of variable Y. We should make it clear that it is not the formulation of TE that is at the origin of mis-attributing the sources of the transferred information. Rather, by definition Shannon's mutual information, $I(X : Y) = H(X) + H(Y) - H(X, Y)$ is dyadic, and cannot capture polyadic correlations where more than one variable influences another. Consider for example a similar but *time-independent* process between binary variables X, Y, and Z where $Z = X \oplus Y$. As is well-known, the mutual information between X and Z, and also between Y and Z vanishes: $I(X : Z) = I(Y : Z) = 0$ (this corresponds to the one-time pad, or Vernam cipher [168], a common method of encryption that takes advantage of the fact that $I(X : Y : Z) = -1$). Thus, while the TE formulation aims to capture a directed dependency of information, Shannon information measures the *undirected* (correlational) dependency of two variables only. As a consequence, problems with TE measurements in detecting directed dependencies are unavoidable when using Shannon information, and do not stem from the formulation of transfer entropy [165] or similar measures such *causation entropy* [179] to capture causal relations. Note that methods such as *partial information decomposition* have been proposed to take into account the synergistic influence of a set of variables on the others [204]. However, such higher-order calculations are

more costly (possibly exponentially so) and require significantly more data in order to perform accurate measurements.

Given the observed error in measuring information flow using TE due to logic gates that encrypt, we now set out to examine how well TE measurements capture information flow when the function is implemented with Boolean functions other than XOR. In particular, we examine every first-order Markov process $Z_{t+1} = f(X_t, Y_t)$ where function f is implemented by all 16 possible 2-to-1 binary relations (Figure 4.1A) and quantify the error in information transfer estimate for each of them. Similar to previous examples, the state of variable Z is independent of its past, and inputs X and Y take states 0 and 1 with equal probabilities, i.e., $P(X = 0) = P(X = 1) = P(Y = 0) = P(Y = 1) = 0.5$.

Table 4.1 shows the results of transfer entropy measurements for all possible 2-to-1 logic gates and the error that would occur if TE measures are used to quantify the information flow from inputs to outputs. This error is the sum of misestimations in information flow quantified by pairwise transfer entropies $\mathrm{TE}_{X\to Z}$ and $\mathrm{TE}_{Y\to Z}$. As we discussed before, for the XOR relation the transfer entropies $\mathrm{TE}_{X\to Z} = \mathrm{TE}_{Y\to Z} = 0$, and $H(Z_{t+1}) = 1$ which means that TE misestimates the information flow from inputs X and Y by 1 bit (the XNOR is exactly the same). We find that in all other polyadic relations where both X and Y influence the future state of Z, $\mathrm{TE}_{X\to Z}$ and $\mathrm{TE}_{Y\to Z}$ capture part of the information flow from inputs to outputs, but $\mathrm{TE}_{X\to Z} + \mathrm{TE}_{Y\to Z}$ is less than the entropy of the output Z by 0.19 bits ($\mathrm{TE}_{X\to Z} + \mathrm{TE}_{Y\to Z} = 0.62$, $H(Z) = 0.81$). In the remaining six relations where only one of the inputs or neither of them influences the output, the transfer entropies correctly capture the information flow. The difference between the sum of transfer entropies, $\mathrm{TE}_{X\to Z} + \mathrm{TE}_{Y\to Z}$, and the entropy of the output $H(Z)$ in XOR and XNOR relations, stems from the fact that $I(X:Y:Z) = -1$, the tell-tale sign of encryption. Furthermore, while other polyadic gates do not implement perfect encryption, they still encrypt partially as $I(X:Y:Z) = -0.19$, which we call *obfuscation*. It is this obfuscation that is at the heart of the TE error shown in Table 4.1.

We repeated similar calculations for the case of a feedback loop network where $Z_{t+1} = f(Y_t, Z_t)$

(Figure 4.1B) and function f can be any one of the 16 logic relations shown in Table 4.1. These simple calculations show that in 16 relations including XOR and XNOR, the sum of the transfer entropies, $\mathrm{TE}_{Y\to Z} + I(Z_{t+1} : Z_t)$ (the formulation for transfer entropy of a variable to itself reduces to processed information $I(Z_{t+1} : Z_t)$) is equal to the entropy of the output Z_{t+1} as was shown in Equation (4.7). However, in XOR and XNOR relations transfer entropy incorrectly attributes all the information to one of the input variables and no influence is attributed to the other. Furthermore, in the polyadic relations other than XOR and XNOR, the transfer entropies $\mathrm{TE}_{Y\to Z}$ and $I(Z_{t+1} : Z_t)$ differ in value while variables X and Y equally influence the state of the output Z, which is why the TE error in these relations is 0.19 bits.

Table 4.1: Transfer entropies and information in all possible 2-to-1 binary logic gates with or without feedback. The logic of the gate is determined by the value Z_{t+1} (second column) as a function of the input X_tY_t=(00,01,10,11). $H(Z_{t+1})$ is the Shannon entropy of the output assuming equal probability inputs, $TE_{X\to Z}$ is the transfer entropy from X to Z. In 2-to-1 gates without feedback, transfer entropies $\mathrm{TE}_{X\to Z}$ and $\mathrm{TE}_{Y\to Z}$ reduce to $I(X_t : Z_{t+1})$, and $I(Y_t : Z_{t+1})$, respectively. Similarly, transfer entropy of a process to itself is simply $I(Z_t : Z_{t+1})$ which is the information processed by Z.

			2-to-1 network, $Z = f(X, Y)$			2-to-1 feedback loop, $Z = f(Y, Z)$		
gate	Z_{t+1}	$H(Z_{t+1})$	$\mathrm{TE}_{X\to Z}$	$\mathrm{TE}_{Y\to Z}$	TE error	$\mathrm{TE}_{Y\to Z}$	$I(Z_t : Z_{t+1})$	TE error
ZERO	(0,0,0,0)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AND	(0,0,0,1)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
AND-NOT	(0,0,1,0)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
AND-NOT	(0,1,0,0)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
NOR	(1,0,0,0)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
COPY	(0,0,1,1)	1.0	1.0	0.0	0.0	1.0	0.0	0.0
COPY	(0,1,0,1)	1.0	0.0	1.0	0.0	0.0	1.0	0.0
XOR	(0,1,1,0)	1.0	0.0	0.0	1.0	1.0	0.0	1.0
XNOR	(1,0,0,1)	1.0	0.0	0.0	1.0	1.0	0.0	1.0
NOT	(1,0,1,0)	1.0	0.0	1.0	0.0	0.0	1.0	0.0
NOT	(1,1,0,0)	1.0	1.0	0.0	0.0	1.0	0.0	0.0
OR	(0,1,1,1)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
OR-NOT	(1,0,1,1)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
OR-NOT	(1,1,0,1)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
NAND	(1,1,1,0)	0.81	0.31	0.31	0.19	0.5	0.31	0.19
ONE	(1,1,1,1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Given that pairwise TE measurements (not taking into account higher-order conditional transfer

entropies) only fail to correctly identify the sources of information flow in cryptographic gates and demonstrate partial errors in quantifying information flow in polyadic relations, we now set out to determine how often these relations appear in networks that implement basic cognitive tasks, and how much error is introduced when measuring information flow using transfer entropy. If the total error in transfer entropy measurements of information flow in cognitive networks is significant, an analysis of pairwise directed information among neural components (neurons, voxels, cortical columns, etc.) using this concept is bound to be problematical. If, however, these errors are reasonably low within biological control structures because cryptographic logic is rarely used, then treatments using the TE concept can largely be trusted.

To answer this question, we use a new tool in computational cognitive neuroscience, namely computational models of cognitive processing that can explain task-performance in terms of plausible dynamic components [93]. In particular, we use Darwinian evolution to evolve artificial digital brains (also known as Markov Brains or MBs [69]) that can receive sensory stimuli from the environment, process this information, and take actions in response. (In the following we refer to digital brains as "Brains", while biological brains remain "brains".). We evolve Markov Brains that perform two different cognitive tasks whose circuitry is thoroughly studied in neuroscience: visual motion detection [21], as well as sound localization [130, 149]. Markov Brains have been shown to be a powerful platform that can unravel the information-theoretic correlates of fitness and network structure in neural networks [45, 8, 164, 114, 85]. This computational platform enables us to analyze structure, function, and circuitry of hundreds of evolved digital Brains. As a result, we can obtain statistics on the frequency of different types of relations in evolved circuits (as opposed to studying only a single evolutionary outcome), and further assess how crucial different operators are for each evolved task, by performing knockout experiments in order to measure an operator's contribution to the task. In particular, we first investigate the composition of different types of logic gates in networks evolved for the two cognitive tasks, and then theoretically estimate how accurate transfer entropy measures could be when applied to quantify the pairwise information flow from one neuron to another in such simple cognitive networks. We then use transfer entropy measures as

a statistic to identify information flow between neurons of evolved circuits using the time series of neural recordings obtained from behaving Brains engaged in their task, and evaluate how successful transfer entropy is in detecting this flow. While artificial evolution of control structures ("artificial Brains") is not a substitute for the analysis of information flow in biological brains, this investigation should provide some insights on how accurate (or inaccurate) transfer entropy measures could be.

4.2 Materials and Methods

4.2.1 Markov Brains

Markov Brains (MB) are evolvable networks of binary neurons (they take value 0 for a quiescent neuron, or 1 for a firing neuron) in which neurons are connected via probabilistic or deterministic logic gates (in this work, we constrain MBs to only use 2-to-1 deterministic logic gates). The states of the neurons are updated in a first order Markov process, i.e., the probability distribution of states of the neurons at time step $t + 1$ depends only on the states of neurons at time step t. This does not imply that Markov Brains are memoryless, because the state of one neuron can be stored by repeatedly writing into its own (or another) neuron's state variable [45, 114, 69]. The connectivity and the underlying logic of the MB's neuronal network is encoded in a genome. Thus, we can evolve populations of MBs using a Genetic Algorithm (GA) [127] to perform a variety of cognitive tasks (for a more detailed description of Markov Brain function and implementation see [69]). In the following sections, we describe two fitness functions designed to evolve motion detection and sound localization circuits in MBs.

4.2.2 Motion Detection

The first fitness function is designed in order to evolve MBs that function as a visual motion detection circuit. Reichardt and Hassenstein proposed a circuit model of motion detection that is based on a delay-and-compare scheme [65]. The main idea behind this model is that a moving object is sensed by two adjacent receptors on the retina, at two different time points. Figure 4.2 shows the schematic of a Reichardt detector in which the τ components delay the stimulus and $\times$ components

multiply the signals, i.e., fires if the signal from the receptor and delay component arrive at the same time. The result of the multiplication units for two different directions is then subtracted so that high values denote motion in one direction (the "preferred direction", PD), low values denote the opposite direction (null direction, ND), and intermediate values encode a stationary stimulus.

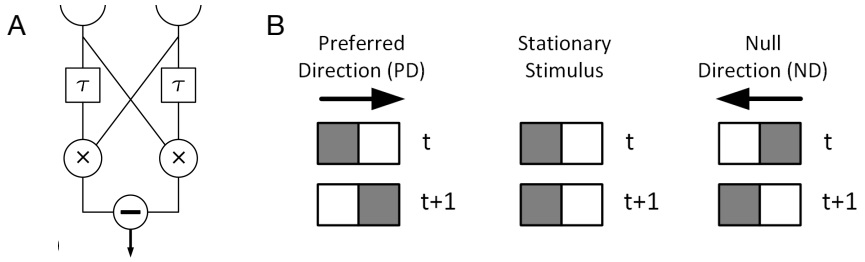

Figure 4.2: (A) A Reichardt detector circuit. In this circuit, the results of the multiplications from each pathway are subtracted to generate the response. The circuit's outcome for PD is +1, ND is -1, and for stationary patterns is 0. (B) Schematic examples of three types of input patterns received by the two sensory neurons at two consecutive time steps. Grey squares show presence of the stimuli in those neurons. The sensory pattern shown here for PD is 10 at time t and 01 at time $t + 1$, which we write as: $10 \rightarrow 01$. Patterns $11 \rightarrow 01$ and $00 \rightarrow 10$ also represent PD. Similarly, pattern $01 \rightarrow 10$ is shown as an example of ND but patterns $11 \rightarrow 10$ and $01 \rightarrow 11$ are also instances of ND.

The experimental setup for the evolution of motion detection circuits is similar to the setup previously used in [184]. In that setup, two sets of inputs are presented to a MB at two consecutive times and the Brain classifies the input as preferred direction (PD), stationary, or null direction (ND). After the first set of inputs, i.e., at time t in Figure 4.2B, a Markov Brain is updated once, and after the second set of inputs (at $t + 1$) it is updated two times, which simulates two operations performed after delaying one of the inputs, namely multiplication and subtraction. The value of the sensory neuron becomes 1 when a stimulus is present, and it becomes 0 otherwise (see Figure 4.2B). Thus, 16 possible sensory patterns can be presented to the MB to classify, among which 3 input patterns are PD, 3 are ND, and the other 10 are stationary patterns. Two neurons are assigned as output neurons of the motion detection circuit. The sum of binary values of these neurons represents the output of the motion detection circuit, 0: ND, 1: stationary stimulus , 2: PD, while in the Reichardt detector circuit shown in Figure 4.2A, the output corresponding to ND is -1, stationary is 0, and PD is +1.

4.2.3 Sound Localization

The second fitness function is designed to evolve MBs that function as a sound localization circuit. Sound localization mechanisms in mammalian auditory systems function based on several cues such as interaural time difference, interaural level difference, etc. [128]. Interaural time difference (which is the main cue behind the sound localization mechanism) is the difference between the times at which sound reaches the two ears. Figure 4.3A shows a simple schematic of a sound localization model proposed by Jeffress [79] in which sound reaches the right ear and left ear at two possibly different times. These stimuli are then delayed in an array of *delay* components and travel to an array of detector neurons (marked with different colors in Figure 4.3A). Each detector only fires if the two signals from different pathways, the left ear pathway (shown at the bottom) and the right ear pathway (shown at top), reach that neuron simultaneously.

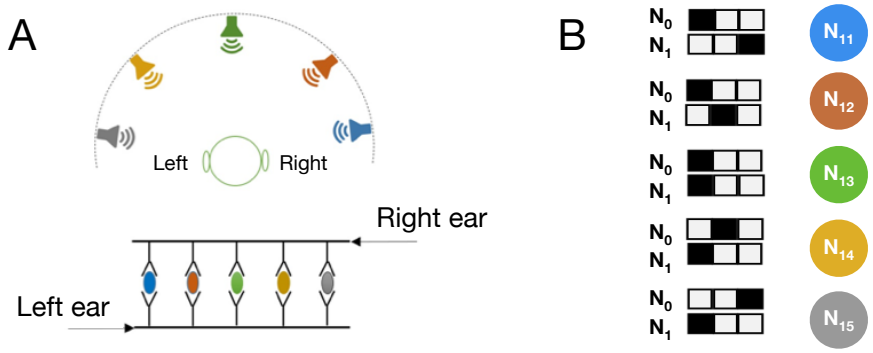

Figure 4.3: (A) Schematic of 5 sound sources at different angles with respect to a listener (top view) and Jeffress model of sound localization. (B) Schematic examples of 5 time sequences of input patterns received by the two sensory neurons (receptors of two ears) at three consecutive time steps. Black squares show presence of the stimuli in those neurons.

In our experimental setup, two sequences of stimuli are presented to two different sensory neurons (neurons N_0 and N_1) that represent the receptors in the two ears. The stimulus in two sequences are lagged or advanced with respect to one another (as shown in Figure 4.3B). The agent receives these sequences and should identify 5 different angles from where that sound is coming. The binary value of the sensory neuron becomes 1 when a stimulus is present, shown as black blocks in Figure 4.3B, and it becomes 0 otherwise, shown as white blocks in Figure 4.3B. Markov Brains are updated once after each time step in the experiment. Similar to the schema shown

in Figure 4.3A, Markov Brains have five designated output neurons (N_{11}-N_{15}) and each neuron corresponds to one of the sound sources placed at a specific angle. Colors of detector neurons (N_{11}-N_{15}) in Figure 4.3B match the angle of each sound source in Figure 4.3A.

4.3 Results

For the motion detection (MD) and sound localization (SL) tasks, we evolved 100 populations each for 10,000 generations, allowing all possible 2-to-1 (deterministic) logic gates as primitives. At the end of each evolutionary run, we isolated one of the genotypes with the highest score from each population to generate a representative circuit.

4.3.1 Gate Composition of Evolved Circuits

Out of 100 populations evolved in motion detection task, 98 led to circuits that perform motion detection with perfect fitness. The number of gates in evolved Brains varies tremendously, with a minimum of four and maximum of 17 (mean=7.92, SD=2.48). The frequency distribution of types of logic gates per each individual Brain is shown for these 98 perfect circuits in Figure 4.4A (in this figure, AND-NOT is an asymmetric AND operation where one of the variables is negated, for example $X' \cdot Y$. Similarly, OR-NOT is an asymmetric OR operation, e.g., $X + Y'$). To gain a better understanding of the distribution of logic gates and how they compose the evolved motion detection circuits, we performed gate-knockout assays on all 98 Brains. We sequentially eliminated each logic gate, (along with all the input and output connections of that gate) and re-measured the mutant Brain's fitness, thus allowing us to estimate which gates were essential to the motion detection function (if there is a drop in mutant Brain's fitness) and which gates were redundant to the motion detection function (if a mutant Brain's fitness remains perfect). The frequency distribution of each type of logic gate per individual Brain for essential gates is shown for the 98 perfect Brains in Figure 4.4B.

For the sound localization task, 71 evolution experiments out of 100 resulted in Markov Brains with perfect fitness. The minimum number of gates was six, with a maximum of 15 (mean=9.14,

SD=1.77). Figure 4.4A shows the frequency distribution of types of logic gates per Brain for these 71 perfect Brains. We also performed a knockout analysis on all 71 evolved sound localization Brains. The frequency distribution of each type of logic gate per individual Brain for essential gates is shown for the 71 perfect Brains in Figure 4.4B. These results demonstrate that the gate type compositions and circuit structures in evolved Brains for motion detection (MD) and sound localization (SL) tasks are significantly different. The total number of logic gates (ignoring duplicates) in the SL task (9.14 gates per Brain, SD=1.77) is greater than the total number of gates in the MD task (7.92 gates per Brain, SD=2.48). Moreover, the number of essential gates in SL (7.13 gates per Brain, SD=1.24) is also greater than the number of essential gates in MD (5.23 gates per Brain, SD=1.31).

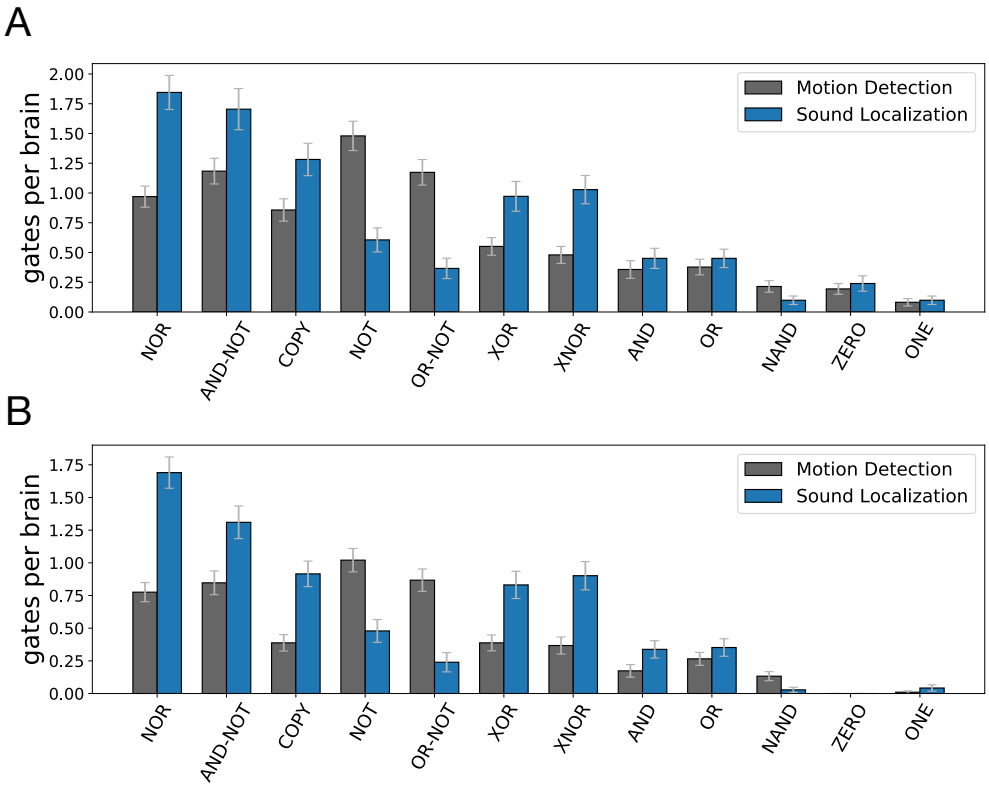

Figure 4.4: Frequency distribution of all, as well as essential, gates in evolved Markov Brains that perform the motion detection or sound localization task perfectly. (A) All gates. (B) Essential gates.

4.3.2 Transfer Entropy Misestimates Caused by Encryption or Polyadicity

As discussed earlier, transfer entropy measures may misestimate the information flow from input to output and may fail to correctly identify the source of information. Table 4.1 gave a detailed analysis of transfer entropy measurements and their misestimates that are rooted either in the polyadic or encrypting nature of the gate, for all possible 2-to-1 logic gates. Given the gate distributions of evolved circuits for motion detection and sound localization tasks along with the misestimate values calculated in Table 4.1, we can estimate the error that would occur when using transfer entropy to quantify the pairwise information flow from source neurons (i.e., input neurons to gates) to receiver neurons (i.e., output neurons of gates). We can similarly estimate what fraction of the information flow from inputs to outputs would be *correctly* quantified by the transfer entropy in the evolved circuits. Recall that in the results presented in Table 4.1, calculations were performed assuming that the input bits take values 0 or 1 with equal probability 0.5. Of course, we cannot generally assume this for the input bits of every logic gate in an evolved network. As a consequence, this analysis only approximates the information flow misestimates of the full network.

In our analysis, we only evaluated the contribution of gates deemed essential via the knockout test. For these essential gates, we summed the pairwise information flow misestimates as well as the correct information flow attributions in each evolved Brain. The mean values of calculated misestimates of information flow as well as correct measurements with their 95% confidence intervals for 98 evolved circuits that perform motion detection task, and for 71 evolved sound localization Brains are shown in Figure 4.5A. In Figure 4.5B, we normalized misestimates and correct measurements by dividing by the number of essential gates in each Brain, and averaged them across Brains. It is worth noting that the calculated information flow misestimates shown in these plots only reflect the misestimates that originated from the polyadicity or encrypting nature of the gates, since they are only based on the network structure and the gate composition of each Brain as well as the analytical results presented in Table 4.1, and do not take into account the errors that could occur as a result of factors such as sampling errors in the dataset or structural complexities in the network, such as recurrent or transitive relations [11, 179, 10]. Along the same

line of reasoning, calculated values of correct measurements represent correct information flows that could be measured by transfer entropy in the absence of the aforementioned sources of errors.

These results further reveal that the circuit structures and gate type compositions in the two tasks are significantly different, and that this structural difference leads to different outcomes when transfer entropy measures are used to detect pairwise information flows. Transfer entropy can potentially capture 3.31 bits (SE = 0.10) of information flow correctly in evolved motion detection circuits (0.64 bits per gate, SE = 0.014), and 3.95 (averaged across 71 Brains, SE = 0.14) bits in evolved sound localization circuits (0.55 bits per gate, SE = 0.014). However, the information flow misestimates when using transfer entropy in evolved sound localization circuits is 2.39 bits (averaged across 71 Brains, SE = 0.12) which is significantly higher than the misestimates in evolved motion detection circuits, which is 1.33 bit (averaged across 98 Brains, SE = 0.085). The information flow misestimate in evolved motion detection circuits is 0.25 bits per gate (SE = 0.014) whereas it is 0.34 bits (SE = 0.016) per gate in evolved sound localization circuits. These findings show that the accuracy of transfer entropy measurements for detecting information flow in digital neural networks can vary significantly from one task to another.

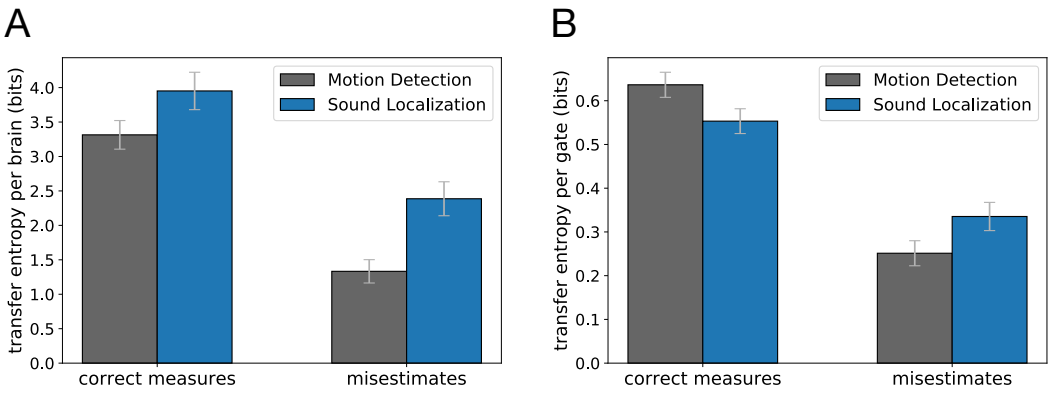

Figure 4.5: Transfer entropy measures, exact measures and misestimates by transfer entropy, on essential gates of perfect circuits for motion detection, and sound localization task. Columns show mean values and 95% confidence interval of misestimates and exact measures (A) per Brain, and (B) per gate.

4.3.3 Transfer Entropy Measurements from Recordings of Evolved Brains

In the previous section we estimated errors in information flow attribution using the error that each particular logic gate in Table 4.1 entails, and then calculated the total error using the gate type distribution for each cognitive task. However as mentioned earlier, this approach only gives a crude estimate of flow because in the evolved cognitive circuits the neurons (and therefore the logic gates) are not independent, and their input is not in general maximum entropy.

Here we use a different approach to assess transfer entropy measurement accuracy in identifying inter-neuronal relations of evolved Markov Brains: we record the neural activities of an evolved Brain when performing a particular cognitive task, similar to the neural recording ("brain mapping") performed on behaving animals. We collect the recordings in all possible trials for each cognitive task and create a dataset for each evolved Brain for that cognitive task. More precisely, for Brains that evolved to perform the motion detection task we record neural firing patterns in 16 different trials. At the beginning of each trial, the Brain is in a state in which all neurons are quiescent. Then, the Brain is updated three times, so we record the Brain's neural activity in 4 consecutive time steps (including the initial state). As a result, the recordings dataset of a Brain that performs motion detection consists of 64 snapshots of the Brain, i.e., the binary state of each neuron. Similarly, a Brain that performs sound localization is recorded during five different trials, and during each trial the Brain is recorded in four consecutive time steps. This results in a recording dataset of size 20 for each evolved Brain. Note that these evolved Brains are deterministic, thus, if a Brain is recorded in the same trial multiple times, its behavior and neural activities remain exactly the same and therefore, recording a Brain once in each trial is sufficient. We then use these recordings to measure transfer entropy for every pair of neurons $\mathrm{TE}_{N_i \to N_j}$ in the network. These transfer entropy measures can be used as a statistic to test whether a neuron N_i causally influences another neuron N_j. Figure 4.6A shows the result of TE calculations performed on neural recording for a Markov Brain evolved in the sound localization task.

To test the accuracy of the TE prediction, we construct an influence map for each neuron of the Markov Brain that shows which other neurons are influenced by a particular neuron. Such a mapping

also determines the receptive field of each neuron, which specifies which other neurons influence a particular neuron. Markov Brains evolve complex networks in which multiple logic gates can write to the same neuron and as a result, it is not straightforward to deduce input-output relations among neurons. Indeed, it was previously argued that even armed with complete knowledge of a given system, finding the causal relation among the components of the system may be a very difficult task [145, 144, 63].

To create our "ground truth" model of direct influence relations, we take into consideration two different components of a Brain's network. First, we take into account the input neurons of a gate and its output neuron, while we also take into consideration the type of the logic gate. For example, in the case of a 'ZERO' gate where the output is always 0 we do not interpret this connection to reflect information flow (as there is no entropy in the output). Second, we analytically extract the binary state of each neuron as a Boolean function of all other neurons using a logic table of the entire Brain (logic table of size 2^{16}, for 16 neurons). This helps us rule out neurons that are connected as inputs to a logic gate while not actually contributing to the output neuron of that gate. Note that this procedure is specifically helpful in cases where more than one logic gate writes into a neuron (when more than one gate writes into a neuron the ultimate result is the bitwise OR of all incoming signals since if either one of them is a non-zero signal it would make the neuron fire, i.e., its state becomes 1). Figure 4.6B shows an example of "ground truth" influence map of neurons for a Brain evolved for sound localization. Each row of this plot shows the influence map of the corresponding neuron and each column represents the receptive field of that neuron. Note that in this plot values are binary, i.e., they are either 0 or 1 which specifies whether a source neuron influences a destination neuron, whereas TE measurements vary in the range [0, 1] bits. Keep in mind that this influence map is only an estimate of information flow gathered from gate logic and connectivity shown in Figure 4.6C.

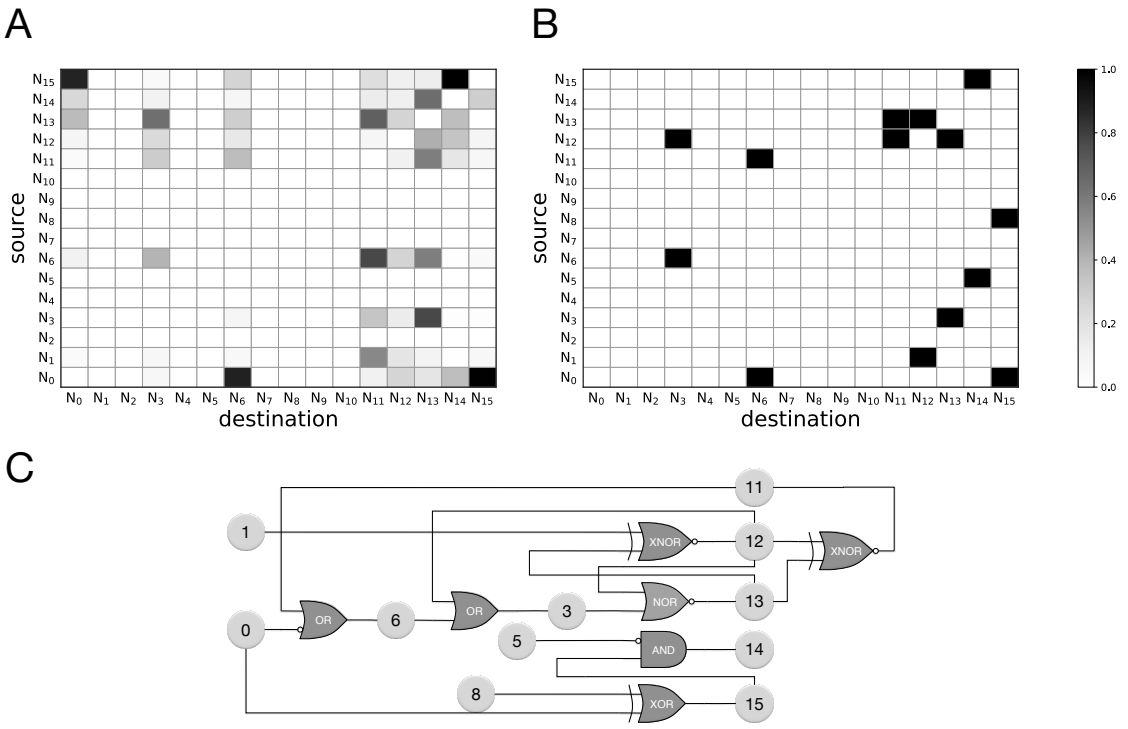

Figure 4.6: (A) Transfer entropy measures from neural recordings of a Markov Brain evolved for sound localization. (B) Influence map (also receptive field) of neurons derived from a combination of the logic gates connections and the Boolean logic functions for the same evolved Markov Brain, shown in (C). (C) The logic circuit of the same evolved Markov Brain; neurons N_0 and N_1 are sensory neurons, and neurons $N_{11} - N_{15}$ are actuator (or decision) neurons.

In order to compare TE measurements with influence maps, we first assume that any non-zero value of the $\mathrm{TE}_{N_i \to N_j}$ implies that there is some flow of information from neuron N_i to N_j. We then evaluate how well TE measurements detect the information flow among neurons based on this assumption. In particular, for each evolved Brain we count 1) the number of existing pairwise information flows between neurons that is correctly detected by TE (hit), 2) the number of relations that are present in the influence map but were not detected by TE (miss), and 3) the number of existing pairwise information flow between the neurons detected by TE measurements that according to the influence map were incorrectly detected (false-alarm). Figures 4.7A and B show the performance results of TE measurements in detecting information flow in Brains evolved in motion detection and sound localization, respectively (averaged across best performing Brains and 95% confidence interval). We observe that the number of false-alarms in motion detection

(mean = 19.0, SE = 0.86) is greater the number of hits (mean = 6.8, SE = 0.20). Similarly, in sound localization the number of false-alarms (mean = 45.1, SE = 1.63) is also greater than the number of hits (mean = 10.1, SE = 0.31), but significantly more so. This again underscores that the accuracy of transfer entropy measures strongly depends on the characteristics of the task that is being solved.

In the results shown in Figure 4.7 we assumed that any value of transfer entropy greater than 0 implies information flow. This assumption can be relaxed such that only transfer entropy values that are greater than a particular threshold imply information flow. We calculated TE measurement performance for a variety of threshold values in the range [0, 1]. The results are presented as receiver operating characteristic (ROC) curves that show hit rates as a function of false-alarm rates as well as their 95% confidence intervals in Figs. 4.7C and D for motion detection and sound localization, respectively [109]. In these plots, the dashed line shows a fitted ROC curve assuming a Gaussian distribution for the $p(\text{TE}|\,\text{information flow is present})$ and $p(\text{TE}|\,\text{information flow is not present})$. The resulting ROC function is $f(x) = \frac{1}{2} erfc(\frac{\mu_1 - \mu_2}{\sqrt{2}\sigma_2} + \frac{\sigma_1}{\sigma_2} erfc^{-1}(2x))$, where $erfc$ is the "error function" complement and $erfc^{-1}$ is the inverse of the error function complement.

In the ROC plots, the datapoint with the highest hit rate (right-most data point) is the normalized result shown in Figure 4.7A, B, that is, the analysis with a vanishing threshold. Note also that the data in Figure 4.7 represent hit rates against false-alarm rate for thresholds spanning the entire range [0,1], implying that hit rates cannot be increased any further unless we assume there is an information flow between every pair of neurons (hit rate=false-alarm rate=1). The false-alarm rates in the ROC curves are actually fairly low in spite of the significant number of false alarms we see in Figure 4.7A, B. This is due to the fact that the number of existing pairwise information flows in a Brain network is much smaller than the number of non-existing flows between any pair of neurons (the influence map matrices are sparse). Thus, when dividing the number of false-alarms by the total number of non-existing information flows, the false-alarm rate is low.

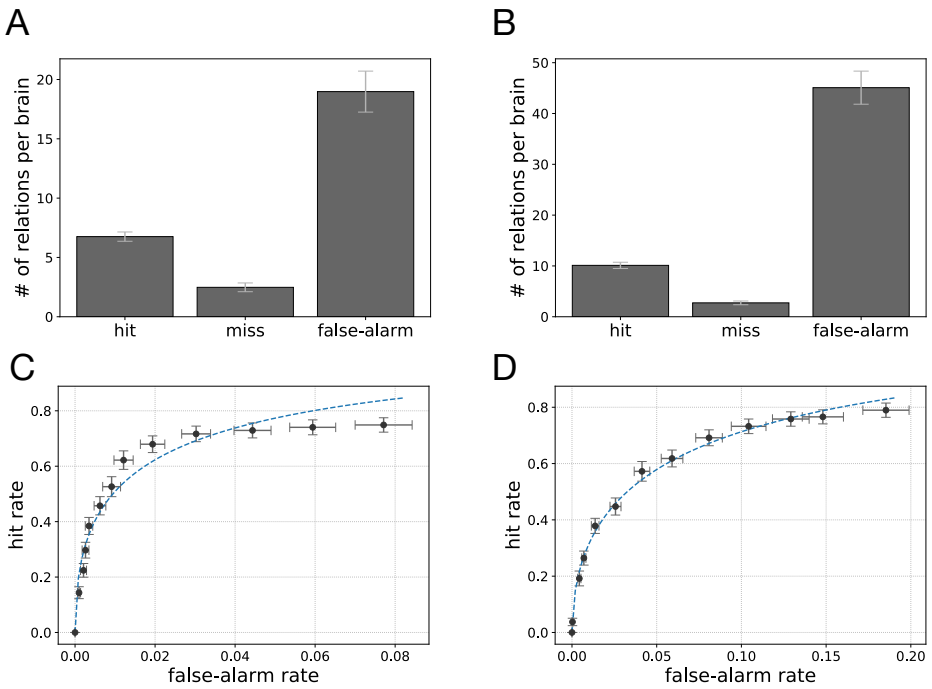

Figure 4.7: Transfer entropy performance in detecting relations among neurons of evolved (A) motion detection circuits, (B) sound localization circuits. Presented values are averaged across best performing Brains along with 95% confidence intervals. Receiver operating characteristic (ROC) curve representing TE performance with different thresholds to detect neurons relations in evolved (C) motion detection, (D) sound localization circuits.

4.4 Discussion

We used an agent-based evolutionary platform to create digital Brains so as to quantitatively evaluate the accuracy of transfer entropy measurements as a proxy for measuring information flow. To this end, we measured the frequency and significance of cryptographic and polyadic 2-to-1 logic gates in evolved digital Brains that perform two fundamental and well-studied cognitive tasks: visual motion detection and sound localization. We evolved 100 populations for each of the cognitive tasks and analyzed the Brain with the highest fitness at the end of each run. Markov Brains evolved a variety of neural architectures that vary in number of neurons and the number of logic gates, as well as the type of logic gates to perform each of the cognitive tasks. In fact, both modeling [152] and empirical [56] studies have shown that a wide variety of internal parameters in neural circuits can result in the same functionality [111]. Thus, it would be informative and

perhaps necessary to examine a variety of circuits that perform the same cognitive task [184].

An analysis of the evolved Brains suggests that selecting for different cognitive tasks leads to significantly different gate-type distributions. Using the error estimate for each particular gate due to encryption or polyadicity, we used the gate-type distributions for each cognitive task to estimate the total error in information flow stemming from using transfer entropy as a statistic. The transfer entropy misestimate was 1.33 bits (SE = 0.08) per Brain on average for Brains evolved for motion detection, whereas in evolved Brains performing sound localization the misestimate was significantly higher: 2.39 bits (SE = 0.12) per Brain on average. More importantly, the inherent differences between the two tasks result in different levels of accuracy when using transfer entropy measures to identify information flow between neurons. It is important to note that in calculating these misestimates, we only accounted for the misestimates that result from TE measurements in polyadic or cryptographic gates. However, we commonly face several other challenges when applying the transfer entropy concept to components of nervous systems (neurons, voxels, etc.). These challenges range from intrinsic noise in neurons to inaccessibility of recording data for larger populations of neurons, which we discuss in more detail later.

We also tested how well transfer entropy can identify the existence of information flow between any pair of neurons using the statistics of neural recordings at two subsequent time points only. Because a perfect model for the "ground truth" of information flow is difficult (if not impossible) to establish, we use an approximate ground truth that uses the connectivity of the network, along with information from the (simplified) logic function to provide a comparison. We find that TE captures many of the connections established by the ground truth model, with a true positive rate (hit rate) of 73.1% for motion detection and 78.7% for sound localization (assuming any non-zero value of transfer entropy implies information flow). The TE measurements miss some relations from the established ground truth while also providing demonstrably false positives, with a false-alarm rate of 7.7% in motion detection and 18.5% for sound localization. For example, some of the information flow estimates in Figure 4.6 manifestly reverse the actual information flow, suggesting a backwards flow that is causally impossible. Such erroneous backwards influence is possible,

for example, when the signal has a periodicity that creates accidental correlations with significant frequency. Besides these false positives, the false negatives (missed inferences) are due to the use of information-hiding (cryptographic or obfuscating) relations, as discussed earlier.

It is noteworthy that in the transfer entropy measurements we performed, we benefited from multiple factors that are commonly great challenges in TE analysis of biological neural recordings. First, our TE measurement results were obtained using error-free recordings of noise-free neurons, while biological neurons are intrinsically noisy. We were also able to use the recordings from every neuron in the network, which presumably results in more accurate estimates. In contrast, in biological networks we only have the capacity to record from a finite number of neurons which, in turn, constrains our understanding of how information flows in the network.

Furthermore, by focusing only on information flow from one time step to the next we can evade the complex issues posed by estimating causal influence, which requires finding optimal time delays in transfer entropies. For example, while a signal may influence a neuron's firing three time steps after it was perceived by a sensory neuron, it must be possible to follow this influence step-by-step in a first-order Markov process, as causal signals must be relayed physically (no action-at-a-distance). As a consequence, when using transfer entropy to detect and follow information flow, we can restrict ourselves to history lengths of 1 ($k = \ell = 1$), which significantly simplifies the analysis [107]. Furthermore, complications arising from discretizing continuous signals [199] do not arise, nor is there a choice in embedding the signal as all our neurons have discrete states. In principle, extending the history lengths (from $k = \ell = 1$ to higher) may be used to reduce false positives in entropy estimates (even for a first-order Markov process), for the simple reason that the higher dimensionality of state space reduces accidental correlations, given a finite sample set. However, such an increase in dimensionality has a drawback: it makes the detection of true positives more difficult (it increases the rate of false negatives) unless the dataset size is also increased. In many dynamical systems such an increase in data size is not an issue, but it may be very difficult (if not impossible) for smaller systems such as the simple cognitive circuits that we evolve. For those, the number of different "sensory experiences" is extremely limited, and increasing the dataset size

does not solve the problem because it would simply repeat the same data. In other words, unlike for large probabilistic systems where generating longer time series will almost invariably exhaustively sample the probability space, this is not the case for motion detection and sound localization. For such "small" systems, increasing the history lengths reduces false positives, but increases false negatives at the same time.

Finally, in order to precisely calculate transfer entropy from Equation (4.1), the summation should be performed over all possible states of variables X_t, Y_t, Y_{t+1}. Using only a subset of those states when calculating the entropy estimate may result in false positives, as well as false negatives. This is another common source of inaccuracy in TE measurements of neural recordings. Here we were able to generate neural recording data for all possible sensory input patterns and included them in our dataset, yet we still observe the described shortcomings in our results. This brings up another important point to notice, namely, even if we introduce every possible sensory pattern to the network, we do not necessarily observe every possible neural firing pattern in the network, and as a result, we do not necessarily sample the entire set of variable states (Y_{t+1}, Y_t, X_t).

4.5 Conclusions

Our results imply that using pairwise transfer entropy has its limitations in accurately estimating the information flow, and its accuracy may depend on the type of network or cognitive task it is applied to, as well as the type of data that is used to construct the measure. Higher-order conditional transfer entropies or more sophisticated measures such as partial information decomposition [204] may be able to alleviate those errors, at the expense of significant computational investments. We also find that simple networks that respond to a low-dimensional set of stimuli (such as the two example tasks investigated here) lead to problems in inferring information flow simply because transfer entropy estimates will be prone to sampling errors.

These findings highlight the importance of understanding the frequency and types of fundamental processes and relations in biological nervous systems. For example, one approach would be to examine transfer entropy in known systems, especially in simple biological neural networks in

order to shed light on the strengths and deficiencies of current methods. Performing an information flow analysis on brains in vivo will remain a daunting task for the foreseeable future, but advances in the evolution of digital cognitive systems may allow us a glimpse of the circuits in biological brains, and perhaps guide the development of other measures of information flow.

CHAPTER 5

MECHANISM OF DURATION PERCEPTION IN ARTIFICIAL BRAINS SUGGESTS NEW MODEL OF ATTENTIONAL ENTRAINMENT

5.1 Introduction

Our ability to deduce causation, to predict, infer, and forecast, are all linked to our perception of time. This activity of the brain refers to an inductive process that integrates information about the past and present to calculate the most likely future event [29]. Without a doubt, this ability is key to an organism equipped with such a brain to survive and prosper, by predicting and deciphering events in the world [134, 160]—as well as the actions of other such organisms. A typical experimental procedure in the study of time perception is comparative duration judgement, in which subjects are asked to compare and judge the duration of events. Generally, duration judgements display the *scalar property*, which implies that the probability distribution of judgements is scale invariant [53]. However, we do not perceive time objectively. Rather, the experience of temporal signals is highly subjective, and is influenced by non-temporal perception, attention, as well as memory [26, 118]. An example of non-temporal perception is the saliency of a stimulus (how it stands out over a background), which may affect how it is perceived.

Attention is another variable that can shape time perception [193, 37, 32, 108, 188]. Because our cognitive bandwidth is limited, we cannot pay attention to all sources of information equally [132]. Rather, a sophisticated mechanism selects which stimuli are attended to, and how much attention is allocated to them. A central hypothesis is that the more attention is devoted to the duration of an event, the longer it is perceived to last [193, 37, 32, 108, 188]. Proposed models of time perception such as Scalar Expectancy Theory (SET) [54] that support this hypothesis usually assume that duration perception is performed with some sort of internal clock [53, 54, 191]. In that model, the onset of an event triggers a switch that starts measuring the accumulation of pulses generated by a *pacemaker*, and triggers the stop switch at the end of the event. The effective rate of pulse

accumulation, in turn, is modulated by the attention given to the stimulus.

In SET, the amount of attention allocated to the stimulus is uniformly distributed in time. By contrast, in models such as Dynamic Attending Theory (DAT) [81, 82, 101] the temporal structure of the signal within which the stimulus is embedded may increase or decrease levels of attention in time. In particular, rhythmic backgrounds can *entrain* the brain so that it expects stimuli to occur periodically, and leads to peaks and troughs of attention. Consequently, models of attentional entrainment based on DAT posit that attentional rhythms that are internal to the cognitive architecture are *synchronised* by external rhythms, so that the external stimuli can then lead to an enhanced processing of events that occur precisely when they are expected to occur [120, 122, 123]. Previous studies have provided support for DAT and related entrainment models, for example by showing that events that occur at rhythmically expected time points can be discriminated more easily than those that occur unexpectedly [101, 122, 83, 124, 129]. In a recent study, McAuley and Fromboluti provided additional support for DAT and related entrainment models by studying the role of attentional entrainment on event duration perception [121]. In that work, they used an auditory oddball paradigm in which a deviant tone (oddball) is embedded within a sequence of otherwise identical rhythmic tones (standard tones). Their results demonstrated that manipulations of oddball tone onset can lead to distortions in oddball tone duration perception. In particular, they observed a systematic *underestimation* of the duration of oddball tones that came early with respect to the rhythm of the sequence, and an *overestimation* of oddball duration in trials where oddballs arrived late with respect the rhythm of the sequence.

Interval timing models such as DAT and SET and their computational counterparts usually take a top-down approach by engineering networks of high-level computational components that describe behavioural/psychophysical data in duration perception [53, 81, 82, 44, 117] (see also references in [5, 64, 61]). Some studies have employed more elaborate models that consist of neuron-scale components [28, 86]. Here, we take a bottom-up approach where evolution leads to a population of diverse computational networks (artificial brains) consisting of lower-level components. These brains may differ in their components and possibly in their behaviours (higher level computations).

These modern computational methods have opened a new path towards understanding perception: the recreation, *in silico*, of neural circuitry that implements behaviour similar to human performance. While this capacity is still in its infancy and therefore can only emulate humans on fairly simple tasks (such as attentional entrainment), the usefulness of this tool for a future "experimental robotic psychology" [19, 2] is evident.

In this study, we use Darwinian evolution to create artificial digital brains, (also known as Markov Brains [69], see Methods), that are able to perform duration judgements in auditory oddball paradigms[1]. Markov Brains are networks of variables with discrete states that undergo transitions evoked by sensory, behavioural, or internal states, and capable of stochastic decisions. As such, they are abstract representations of micro-circuit cortical models [62], except that their dynamics is not programmed.

We run 50 replicates of the evolutionary experiment (i.e., 50 different populations) and from each pick the best-performing Brain. These evolved Brains display behavioural characteristics that are similar to human subjects: for example, their discrimination thresholds satisfy Weber's law. In fact, these 50 Brains can be thought of as participants in a cognitive experiment. We then test these Brains against auditory oddball paradigms that they have never experienced before, in which the oddball tone may come early or late with respect to the rhythm of the sequence (similar to the first series of experiments in Ref. [121]). The evolved Brains show distortions in perception of early/late tones similar to what was reported in human subjects [121]. We then analyse the algorithms and computations involved in duration judgement in order to discover how these algorithms result in systematic distortions in perception of early/late oddballs.

Our findings demonstrate that the computations involved in duration judgements and distortions is quite different from existing time perception theories such as scalar expectancy theory (SET) or dynamic attending theory (DAT), and suggest a new theory of perception in which attention to uncertain parts of the stimuli plays the central role, whereas predictable parts require less attention (i.e., less processing) because they are expected [78]. This is consistent with recent findings that

[1]Here and below, to avoid confusion we use "Brain" with a capital B to denote artificial brains, while biological brains remain just "brains".

predictability of stimuli results in more rapid recognition [119]. We close with speculations that suggest a broader view in which all cognitive processing can be understood in terms of context-dependent prediction algorithms that pay attention only to those parts of the signal that are predicted to have the highest uncertainty, and are therefore likely to be informative.

5.2 Results

We evolve Markov Brains that are capable of duration judgements of an oddball tone placed in a rhythmic sequence of identical tones (standard tones) with a variety of standard tone durations and inter-onset-intervals (IOI) (Fig. 5.1 shows a schematic of the auditory oddball paradigm). We ran 50 replicates of the evolution experiment for 2,000 generations and from each population picked the Brain with the highest performance at the end of each run. The best performing Brains in all 50 populations gain 98.0% fitness on average (see Fig. 5.10).

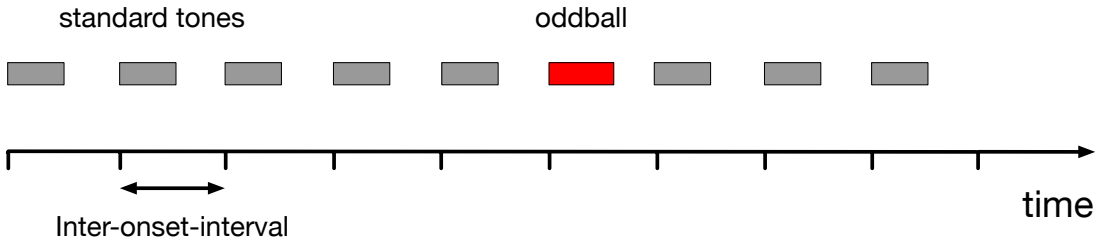

Figure 5.1: A schematic of the auditory oddball paradigm in which an oddball tone is placed within a rhythmic sequence of tones, i.e., standard tones. Standard tones are shown as grey blocks and the oddball tone is shown as a red block. Oddball tone duration may be longer or shorter than the standard tones.

5.2.1 Discrimination thresholds of evolved Markov Brains comply with Weber's law

We used average responses of the evolved Brains to generate psychometric curves as follows. For each (IOI, standard tone) we averaged the decision responses of 50 evolved Brains. Using these averaged responses, we generated psychometric curves corresponding to each standard tone as prescribed by [109] and calculated the point of subjective equality (PSE) and just noticeable difference (JND). The PSE measures the duration for which Markov Brains respond longer (or

shorter) 50% of the time which, in essence, marks the duration of the oddball that is perceived to be equal to the standard tone. The JND measures the sensitivity of the discrimination, or discrimination threshold, for a standard tone. In other words, the JND represents the slope of the psychometric curve, where steeper slopes show higher discrimination sensitivity or lower discrimination threshold. The PSE reflects the accuracy of the perception while the JND indicates its precision. The values of PSE, JND, and their standard deviations are presented for all inter-onset-intervals and standard tones in Table 5.1.

Table 5.1: This table contains point of subjective equality (PSE), just noticeable difference (JND), and their standard deviations (SD), as well as relative JNDs, and constant error (CE) of on-time oddballs for all inter-onset-intervals, standard tones. Responses are averaged across all 50 Brains to generate psychometric curves.

IOI, std tone	PSE	PSE SD	JND	JND SD	relative JND	CE
(10, 5)	4.89	0.073	0.335	0.050	0.067	-0.109
(11, 5)	5.09	0.068	0.292	0.039	0.058	0.092
(12, 6)	5.92	0.083	0.458	0.051	0.076	-0.077
(13, 6)	6.27	0.072	0.339	0.050	0.057	0.265
(14, 7)	6.80	0.080	0.404	0.050	0.058	-0.204
(15, 7)	7.05	0.072	0.416	0.041	0.059	0.049
(16, 8)	7.76	0.067	0.380	0.037	0.047	-0.242
(17, 8)	8.05	0.064	0.402	0.038	0.050	0.051
(18, 9)	8.58	0.072	0.372	0.049	0.041	-0.417
(19, 9)	9.01	0.081	0.469	0.048	0.052	0.012
(20, 10)	9.76	0.078	0.403	0.052	0.040	-0.240
(21, 10)	10.45	0.093	0.442	0.045	0.044	0.448
(22, 11)	11.19	0.109	0.655	0.085	0.060	0.192
(23, 11)	11.99	0.116	0.756	0.075	0.069	0.993
(24, 12)	13.04	0.119	0.829	0.071	0.069	1.036
(25, 12)	13.82	0.128	0.900	0.079	0.075	1.819

According to Weber's law [47], the discrimination threshold (e.g., the JND) varies in proportion to the standard stimulus; therefore, the values of relative JND, defined as $\frac{\text{JND}}{\text{std tone}}$, should remain constant. Getty showed that empirical results of duration perception in the range of 80 msec to 2 seconds is explained very well with Weber's law [52]. Fig. 5.2A shows the psychometric curves generated from the averaged responses of all 50 Brains for every (IOI, standard tone). In this figure, durations are normalised by standard tone. Psychometric curves for different standard

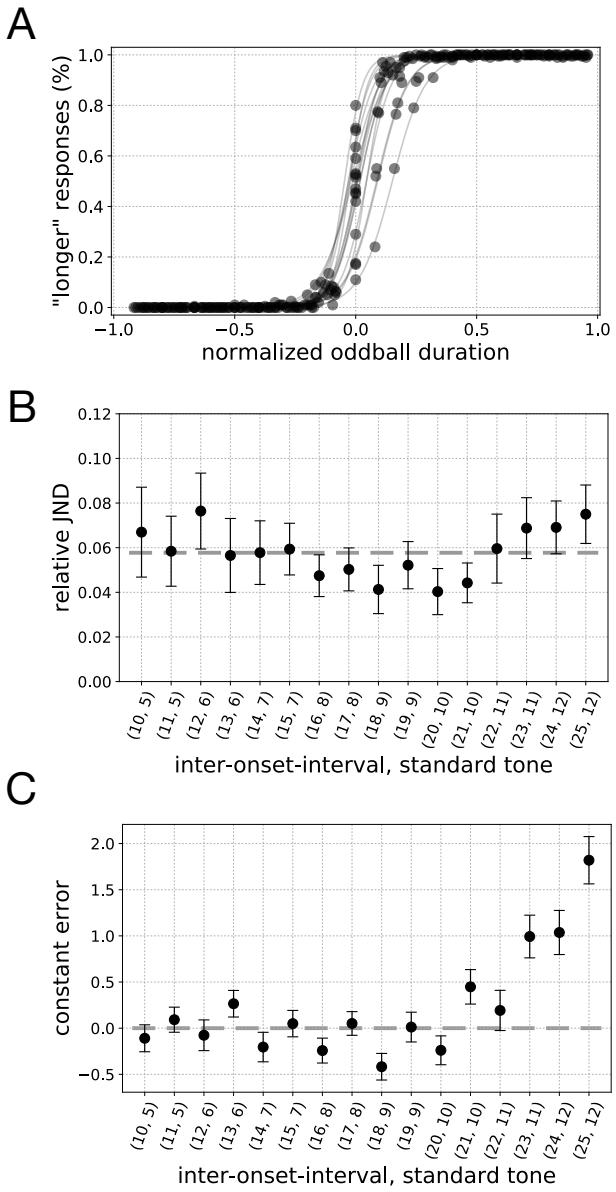

Figure 5.2: (A) Psychometric curves generated from averaged responses of 50 evolved Brains for every inter-onset-interval, standard tone. Oddball durations on the x-axis are normalised by standard tone to lie in the range (-1, 1). (B) Relative JND values and their 95% confidence interval as a function of inter-onset-interval, standard tone. Dashed line shows the average value of relative JNDs. (C) Constant errors, the difference between PSE and standard tone, and their 95% confidence interval as a function of inter-onset-interval, standard tone. Dashed line shows CE=0.

tones overlap, which shows that relative JNDs in all these trials are similar and confirms that they are in accordance with Weber's law. Fig. 5.2B shows relative JNDs as a function of standard tones. All relative JNDs are in the range between 0.04 and 0.07 with mean=0.06 and standard

deviation=0.01, similar to the values found in [52]. The difference between PSE and the standard tone, also known as the constant error (CE), shows the deviation of perceived duration of tone from its actual duration. The values of CE are shown for every (IOI, standard tone) in Fig. 5.2C and we observe that for longer IOIs, CE values start to deviate slightly from zero. This deviation in PSE values for longer tone durations is also observed in human subjects [52]. However, this deviation of CEs for longer tones in Markov Brains was different from human subjects in that CE values in human subjects start decreasing for longer durations (they are negative) whereas in Markov Brains CE values increase (they are positive). This difference can be attributed to the fact that in the experiments described in Ref. [52] subjects do not receive any feedback about their performance duration judgements whereas Darwinian evolution provides feedback implicitly via selection. The mechanisms behind the distortion in duration perception in longer IOIs are explained in more detail in Additional Experiments and Analysis.

5.2.2 Evolved Brains show systematic duration perception distortion patterns similar to human subjects

In the next step, we tested evolved Markov Brains with stimuli that they had never experienced during evolution, namely oddballs that arrive early or late with respect to the rhythm of the sequence of tones (termed "test trials"). In trials used during evolution ("training trials"), oddballs always occurred in sync with the rhythmic tone (on-time oddballs). These test trials included all possible oddball durations but also all possible oddball onsets, meaning oddballs were delayed or advanced as many time steps as possible as long as they did not interfere with the following or preceding tone. Then, we used the average response of 50 Brains to generate psychometric curves for early/late oddballs, and to calculate PSE values.

We used PSE values to calculate the duration distortion factor (DDF), defined as the ratio of the point of objective equality (the standard tone) and the point of subjective equality (PSE). Fig. 5.3 shows the DDF as a function of the onset of the oddball for all IOIs. In this plot, negative onset values stand for early oddballs and positive values of onset represent late oddballs. A DDF

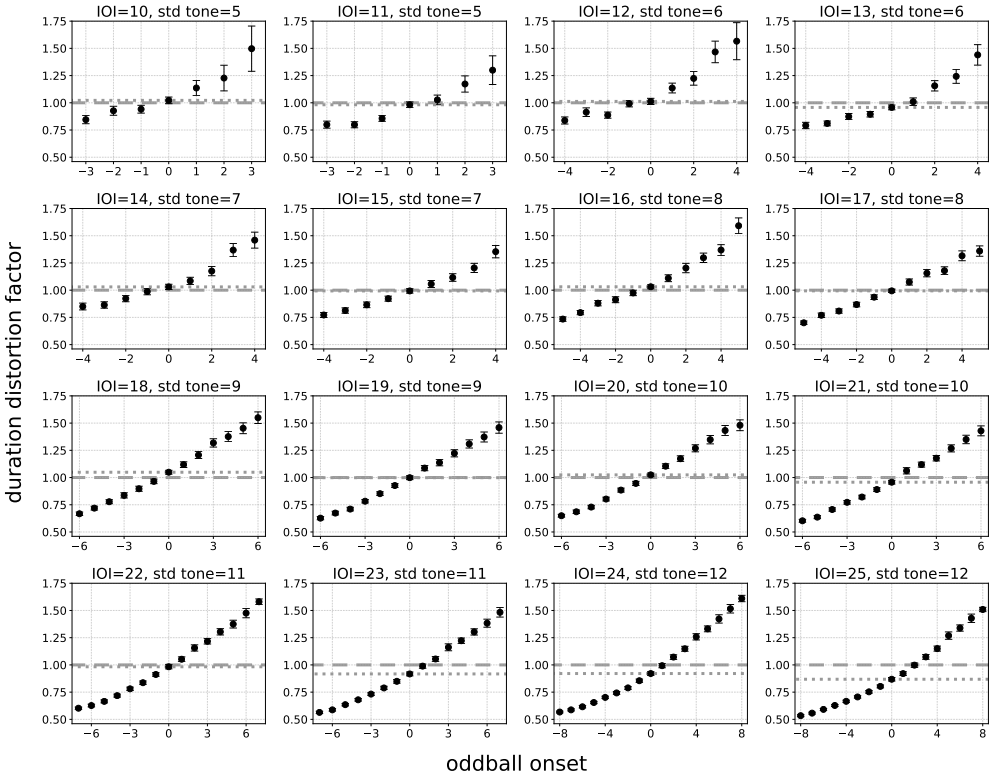

Figure 5.3: Duration distortion factors (DDF) and their 95% confidence interval as a function of the onset of the oddball for all IOI, standard tones. Negative onset values represent early oddballs and positive values of onset represent late oddballs. A DDF greater than 1 shows an overestimation of the duration of the oddball and DDF less than unity shows an underestimation of the duration of the oddball. The dashed line indicates DDF=1 and the dotted line shows DDF for on-time oddball tone.

greater than one shows an overestimation of the duration of the oddball whereas a value less than unity reflects an underestimation of the duration of the oddball. Just as was observed with human subjects [121], the late oddballs are perceived as longer and the early oddballs are perceived as shorter compared to the standard tone. In addition, the more delayed (early) the oddball tone, the more its duration is overestimated (underestimated) compared to the standard tone, which is again consistent with results presented in experiment 2 of Ref. [121].

5.2.3 Algorithmic analysis of duration judgement task in Markov Brains

The logic circuits of evolved Markov Brains are complicated and defy analysis in terms of causal logic. As observed before, these networks turn out to be "epistemologically opaque" [116], in the sense that their evolved logic does not easily fit into the common logical narratives we are familiar with. Rather than focus on the Boolean logic of Markov Brains, we here focus on their state space [55, 166]. In particular, we investigate the state transitions and how these transitions unfold in time, in order to discover the computations that are at the basis of the observed behaviour [16].

5.2.3.1 Temporal information about stimuli is encoded in sequences of Markov Brain states

Evolved Brains display periodic neural activation patterns in response to rhythmic auditory signals (this is, by definition, entrainment). These periodic neural firing patterns translate to loops in state transition diagrams (see Methods for more details on state transitions in Markov Brains). In each trial, the first few tones an evolved Brain listens to typically shift the Brain's activation pattern towards a region in state space that is associated with this rhythm. More precisely, the opening tones transition the Brain to a sequence of states that form a loop in the state-to-state diagram, and the Brain remains in that loop as long as the stimulus is repeated. Fig. 5.4A shows an instance of a Markov Brain state transition diagram when listening to rhythmic tones with IOI=10 and standard tone=5 in the absence of an oddball. The state of the Brain is calculated from equation (5.2). Supplementary Movie 1 shows the state-to-state transitions as the Brain listens to a sequence of standard tones. This sequence of Brain states encodes the contextual information about the stimuli, that is, the sequence forms an internal representation of the rhythm and the standard tone. More importantly, this sequence produces an expectation of future inputs that enables the Brain to compare the input it has sensed with future inputs. In particular, when the Brain receives the oddball, it usually transitions out of this loop to follow a different trajectory in state space (see for example Fig. 5.4B) to judge the oddball duration, which is a comparison mechanism between the standard tone (what is expected) and the oddball. Fig. 5.5 shows that in most of the trials (77.6% of the trials) Brains evolve loops of the same size as the period of the rhythmic tones (the IOI), but

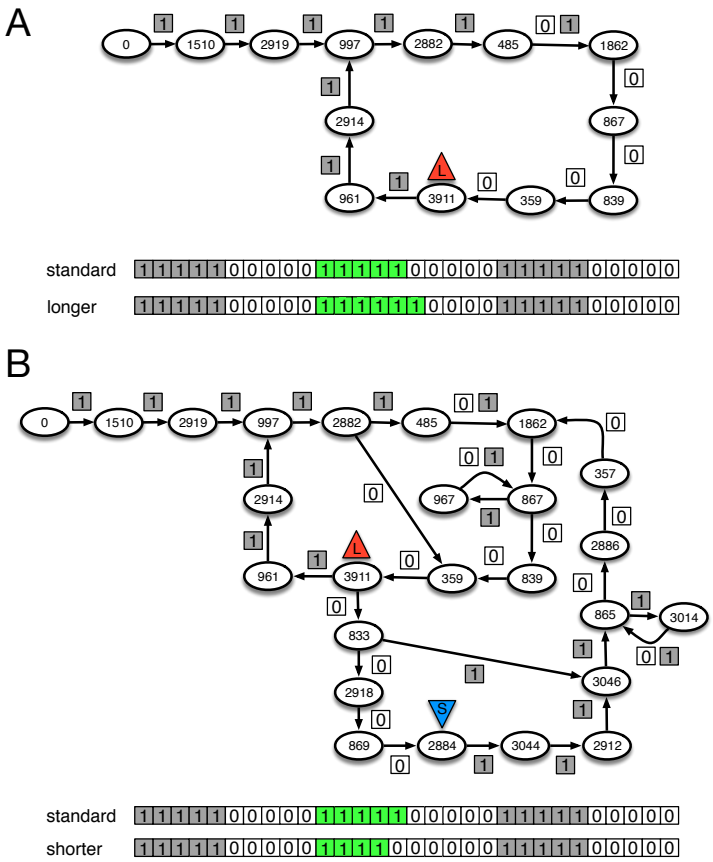

Figure 5.4: State-to-state transition diagram of a Markov Brain for IOI=10, and standard tone=5, with oddball tones of duration 5, 6 shown in (A) and 4 shown in (B). Before the stimulus starts, all neurons in the Brain are quiescent so the initial state of the Brain is 0. The stimulus presented to the Brain is a sequence of ones (representing the tone) followed by a sequence of zeros (denoting the intermediate silence). The stimulus at each time step is shown as the label of the transition arrow in the directed graph. The input sequence is shown for the standard and oddball sequences at the bottom of the state-to-state diagrams. (A) State-to-state transition diagram of a Markov Brain when exposed to a standard tone of length 5, as well as a longer oddball tone of length 6. This Brain judges an oddball tone of duration 6 by following the same sequence of states as the original loop, because the transition from state 485 to 1862 occurs irrespective of the sensory input value, 0 or 1. This Brain correctly issues the judgement "longer" from state 3911, indicated by the red triangle at the end of the time interval (see Supplementary Movie 1 and Supplementary Movie 2 for standard tone and longer oddball tone, respectively). (B) The state-to-state transition diagram of the same Brain when presented with a shorter oddball tone of length 4. The decision state is marked with a down-pointing blue triangle. Once the Brain is entrained to the rhythm of the stimulus, the shorter oddball throws the Brain out of this loop. The exit from the loop transitions this Brain into a different path. After four ones the Brain transitions to state 359 (instead of continuing to 485), and then continues along a path where it correctly judges the stimulus to be "shorter" in state 2884 (see also Supplementary Movie 3).

some Brains have loops that are multiples of the IOI. In this figure, the size of the each marker is proportional to the number of Brains that evolve a particular loop length in each IOI. Also, further analysis shows that in 93.6% of trials, evolved Brains transition out of these loops at the exact time point where there is a mismatch in oddball and standard tone.

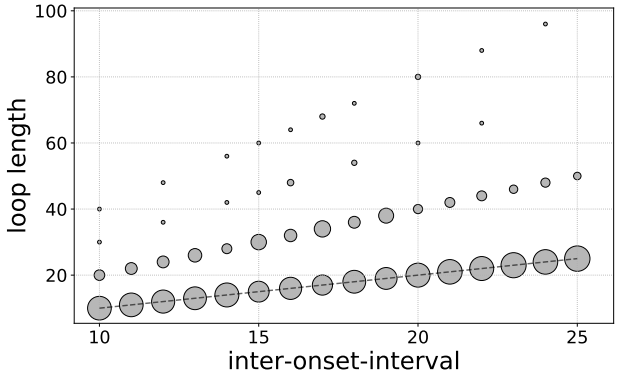

Figure 5.5: The distribution of loop sizes of 50 evolved Brain for each inter-onset-interval (IOI). The size of the markers is proportional to the number of Brains (out of 50) that evolve a particular loop length in each IOI. The dashed line shows the identity function.

5.2.4 Algorithmic analysis of distortions in duration judgements: Experience and perception during misjudgements of early/late oddballs

The similarity of behavioural characteristics in the perception of event duration between Markov Brains and human subjects appears to imply a fundamental similarity between the underlying computations and algorithms. In the following, we present brief definitions of concepts such as attention, experience, and perception in terms of state transitions in deterministic finite-state machines that are later used in our analysis (in Methods we present more formal definitions of these concepts and the reasoning behind them).

1) Attention to a stimulus: When a Brain is in state S_t and transitions to state S_{t+1} regardless of the stimulus (zero or one), we say the Brain *does not pay attention* to the input stimulus. More generally, a Brain pays less attention to an input stimulus or a sequence of stimuli if that input does not affect the state of the Brain later in the future state, S_{t+k}.

2) Perception of a trial: The state of the Brain at the end of the oddball tone interval (when it issues the longer/shorter decision) is the Brain's perception of the tone sequence.

3) Experience of the stimuli: The temporal sequence of Brain states when exposed to a sequence of input stimuli constitutes the Brain's experience.

We first hypothesised that early or late oddball tones drive the Brain into states that they had never visited before (as these Brains had never previously experienced early or late oddball tones) and that these new states are responsible for misjudgements of early or late oddballs. When exposed to late or early oddballs, Brains visited on average 22.26 (SE=4.33) new states across 50 evolved Brains, approximately 32% of the number of states they visited during trials with on-time oddballs, which is 69.80 states on average (SE=5.07). We then tested how often these new states are decision states for the misjudgements of out-of-time oddball tones. Our tests show that in such misjudgements, the Brain state at decision time point is almost *never* a new state that has not appeared before (it happened in one test trial for one Brain out of 56,250 different test trials in all 50 Brains).

Given that during misjudgements of out-of-rhythm oddballs the decision state is a state that had previously occurred during evolution, we test whether there is any connection between Brain states during these misjudgements and Brain states in training trials. In other words, we investigate how the experience during a misjudgement relates to experiences the Brain had in its evolutionary history. In the next two sections, we address these questions by separately focusing on perception and experience of Markov Brains during misjudgements of out-of-rhythm oddball tones.

5.2.4.1 The onset of the tone does not alter a Brain's perception of the tone

Our null hypothesis is that the perception of an out-of-rhythm oddball tone may be any one of the states that the Brain has traversed in training trials with equal probability. In any of these Brain states, the decision neuron will be either quiescent or firing, so we call the set of states with quiescent decision neuron "shorter-judging states" denoted as S_{Sh}, and the set of states with firing decision neuron "longer-judging states" denoted as S_{Lo}. Thus, the probability that a Brain at

decision time is in any of the shorter-judging states, for example, is calculated by

$$\text{Prob}(S_{\text{decision}} \in S_{Sh}) = \frac{1}{|S_{\text{Sh}}|}, \tag{5.1}$$

where $|S_{\text{Sh}}|$ is the cardinality of the set of shorter-judging states, and similarly, $\text{Prob}(S_{\text{decision}} \in S_{Lo}) = \frac{1}{|S_{\text{Lo}}|}$.

We develop our alternative hypothesis that captures possible associations between experience and perception during misjudgement of out-of-rhythm oddballs and experiences and perceptions they had in training trials. In order to discover such possible associations, for any given misjudgement of early or late oddball we limit our search domain to training trials with the same inter-onset-interval and standard tone as the misjudgement trial. In the next step, we search for correlations between the perception and various oddball tone properties such as its 1) onset (time step at which the oddball begins, T_{init}), 2) duration (ΔT), and 3) ending time point (time point at which the oddball ends, T_{fin}). To this end, we calculated the information shared between the perception and oddball tone properties (see Methods for a detailed explanation of information computation procedures). Fig. 5.6A shows the information shared between the perception (decision state S_{decision}) of the Brains and 1) oddball ending time (shown in grey), 2) oddball onset (shown in blue), and 3) oddball duration for each inter-onset-interval and standard tone. These results show that the oddball ending time point is a better predictor of the perception than the oddball tone onset or its duration. Note also that the information shared between the perception and the oddball ending time point remains consistent across all IOI and standard tones, whereas shared information between perception and oddball duration, and perception and onset decrease monotonically as IOI and standard tones increase. Building on these results, we propose the following alternative hypothesis: during misjudgement of an early or late oddball, a Brain goes through a state sequence that is reminiscent of experiences it had during trials with the same IOI and standard tone, and with on-time oddballs that end at the same time point as the early or late oddball (an example scenario is shown Fig. 5.6B).

In order to test this alternative hypothesis, we perform another test to measure how often perception in misjudgement of early or late oddballs is identical to perceptions in similar training

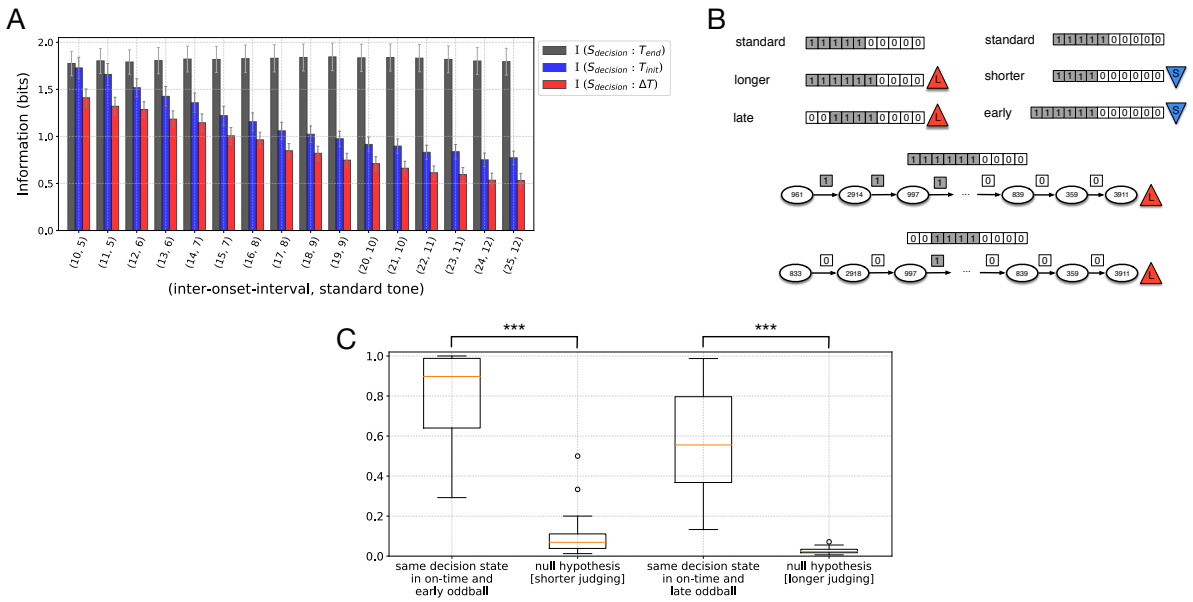

Figure 5.6: (A) The mutual information between perception, i.e., the decision state of the Brain, and 1) the oddball tone ending time step (shown in black), 2) the oddball tone duration (shown in red), 3) the oddball tone onset (shown in blue), and their 95% confidence intervals. (B) Sequence of inputs for a standard tone, an on-time longer oddball tone that is correctly judged as longer, and a shorter late oddball tone that is misjudged as longer. Sequence of inputs for a standard tone, an on-time shorter oddball tone that is correctly judged as shorter, and a longer early oddball tone that is misjudged as shorter. Sequences of Brain states along with input sequences for on-time longer oddballs and shorter late oddballs.(C) The fraction of misperceived out-of-time oddball tones that resulted from having the same perception in on-time and out-of-time stimuli with the same oddball end points (left data point), compared to the null hypothesis; likelihood that Brains misjudgements were to be issued from any one of states from set of "shorter-judging" or "longer-judging" states (middle and right data point, respectively).

trials. Consider for example a trial with IOI=10, standard tone=5 with a late oddball tone (onset=2) that is shorter than the standard tone (duration=4) as shown in Fig. 5.6B. When a Brain misjudges this oddball as "longer" (with $S_{\text{decision}} = 3911$ as shown in Fig. 5.6B), we search for instances in the set of training trials (with on-time oddball) with IOI=10 and standard tone=5, where that Brain issued a correct "longer" decision for an oddball that ended at the same time point as the late shorter oddball (as shown in Fig. 5.6B). The same analysis can be performed for misjudgement of early oddballs that are longer than the standard tone (Fig. 5.6B). We count the number of such instances for each Brain and divide the result by the total number of its misjudgements of out-of-rhythm oddball tones. Fig. 5.6C (left data point) shows the result of this analysis for all 50

Brains. This result shows that in the vast majority of the cases (with median of 69.5% of the cases), the misjudged out-of-rhythm oddball and on-time oddballs that end at the same time point are *perceived* the same. In other words, the misjudgement is due to Brains paying less attention to the onset of the tone, meaning the onset of the oddball does not affect the ultimate state from which the decision is issued. The middle and left data points show the probabilities calculated from equation 5.1 described in the null hypothesis, that measure how likely it is for a Brain to, by chance, end up in any of the "shorter-judging" or "longer-judging" state at decision time. Our statistical analysis shows that having the same decision state in out-of-rhythm oddball and on-time oddballs (with constraints explained above) are significantly more likely than being in "shorter-judging" (median=0.695 vs. median=0.069, Mann Whitney $U = 2494.0, n = 50, p = 5.03 \times 10^{-18}$ one-tailed) or "longer-judging" state at decision time (median=0.695 vs. median=0.023, Mann Whitney $U = 2500.0, n = 50, p = 3.51 \times 10^{-18}$ one-tailed), therefore, we reject the null hypothesis in favour of the alternative hypothesis.

Based on these findings, we conclude that during misjudgements of early or late oddball tones, Markov Brains pay more attention to the end point of the oddball and less attention to the oddball duration, or it onset. This is presumably because during evolution tones are always rhythmic and Brains that entrain to the rhythm expect the oddball to be on-time. As a result, Brains pay more attention to when the oddball ends which is a more informative component of the stimuli than its onset which, during evolution, had no variation and hence, no uncertainty.

5.2.4.2 Experience of early or late oddball is similar to adapting entrainment to phase change

Here we investigate the entire sequence of Brain states (Brain experiences of the stimuli) for those instances we found in previous section, in which the perception of the Brain in misjudgement of early/late oddballs was the same as perception of shorter/longer on-time tones with the same end point as an out-of-time oddball. In order to compare two experiences, we use two different measures (experience comparison is a form of representational similarity analysis, see for example [96, 95]). First, we find the longest common sub-sequence that includes the decision state. In other words, we

start from the decision state in on-time and out-of-time sequences (note that the decision state is the same in both sequences), trace back the transitions in sequences and count the number of states that are identical in both sequences until the first mismatch occurs. The length of the identical portion of the two sequences is then normalised by the total length of one sequence (recall that the length of both sequences are the same) to lie in the range (0,1], we term this normalized length of the identical portion of experiences the *similarity depth*, since it measures how deeply the on-time and out-of-time oddball experiences are identical. We note that because the perception of the tone is the same in these trials, the similarity depth must be greater than zero. Second, we use the Jaccard index, that measures the overall similarity of sequences by comparing states at same positions in the two sequences.

Fig. 5.7A shows the distributions of similarity depth and total similarity of experiences. Fig. 5.7B shows the distribution of the difference between the similarity depth and total similarity. The difference between the two measures is zero in 91.5% of the cases which implies that the experiences are almost always entirely different up to the point where they become identical. We observe a wide variety in these similarity measures which shows that Brains do not traverse the exact same trajectory they did during an on-time trial; rather the early or late oddball initially throws the Brain out of this trajectory but later the Brain returns to states it experienced during an on-time oddball with the same end point. In other words, the onset of the out-of-time oddball is noticed, however, since the Brains are entrained to the rhythm and expect the oddball to be on-time their computations of duration relies more on their expectation than the actual start point of the oddball. This mechanism is reminiscent of adapting to phase changes in entrainment to rhythmic stimuli.

5.3 Discussion

This study was aimed at elucidating the neural (mechanistic) underpinnings of perception, by evolving digital Brains that perform duration judgements of tones that were presented in a rhythmic sequence, and that were later subjected to out-of-rhythm oddball tones to quantify distortions in

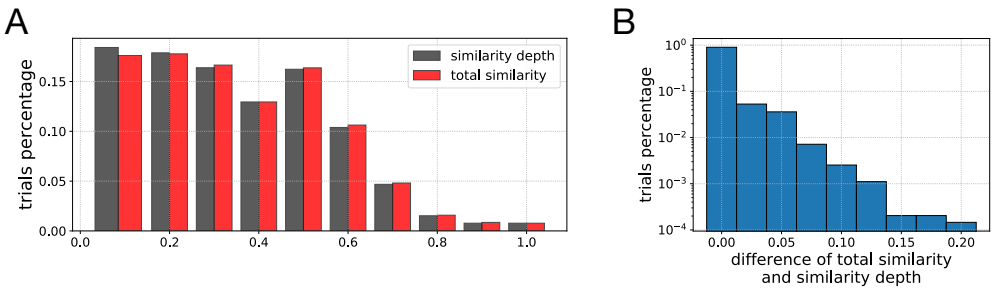

Figure 5.7: (A) Distribution of similarity depth of experiences (sequences of states) of on-time and early/late oddball tones in trials in which onset does not change the perception of the tone in Markov Brains. Similarity depth one implies that the experiences are identical throughout the tone perception. (B) The distribution of the difference between the total similarity and similarity depth in each trial.

duration judgement that occur as a response to the onset manipulation. We found that evolved Markov Brains display a capacity to discriminate tone length that is remarkably similar to people's ability to distinguish changes (quantified by Weber's Law) to the extent that the observed relative JND of Markov Brains was in the same range (6-10%) as in some of the experiments in [52, 121]. Furthermore, evolved Markov Brains exhibit a systematic distortion in perceived event duration of out-of-rhythm oddball tones that is also similar to what was observed in a human subjects study previously conducted by one of the authors. But while the conclusion of [121] was that the experiments supported the dynamic attending theory (DAT) of attentional entrainment (which, we recall, posits that entrainment creates peaks of attention that coincide with the start of each tone) we here find instead that Markov Brains pay attention to the *end of the signal*, and pay less attention to the onset.

From the point of view of Bayesian inference [92], a model of cognition that focuses attention on those parts of the signal that carry most of the uncertainty (the end of the stimulus) makes eminent sense. After all, Brains that have experienced only on-time stimuli should take the rhythmic nature of stimuli for granted: there is no need to pay attention to predictable stimuli. In fact, this view of cognition is fully consistent with the Hierarchical Temporal Memory (HTM) model of neo-cortical computation [67], which is based on the idea that brains are prediction machines. This model of attention differs from common models of visual processing and attention such as visual and

auditory saliency [74, 87], because in those models only the contrast of the stimulus with the background is considered for saliency, not the value of the information it contains. The model is consistent, however, with neurophysical models in which temporal anticipation improves perception but does *not* affect the spontaneous firing rate [78, 119], which is associated with attention in visual processing [192].

The present work suggests a model of cognition where the stimulus not only entrains the cognitive apparatus, but conditions the brain to expect only a small subset of possible future states. From this point of view, any temporal history of stimuli leads to predictions that, for the most time, will come to pass unless the environment has changed in a way that necessitates further attention. In particular, our findings suggest that both DAT and SET are incomplete models of time perception where DAT unduly emphasises attention peaks at the beginning of each tone in the sequence, while SET uses the onset and the end of the tone to start and stop a clock, contrary to our (admittedly digital) evidence.

The results presented here open up a number of different questions and avenues for future exploration. Can the theory of dynamical entrainment we present here be meaningfully tested in human experiments, by focusing on those predictions that distinguish it from established theories such as SET and DAT? Does this theory also explain observations in different sensory modalities such as vision? A program in which empirical studies using human subjects coupled with sophisticated digital experimentation might provide an answer, and open up avenues for a detailed mechanistic understanding of the complexities of perception. Ultimately, this opens up the possibility of explaining phenomenological concepts such as attention, perception, and memory in terms of state-space dynamics of cortical networks.

5.4 Methods

The use of mathematical and computational methods for the study of behaviour is growing, especially due to the unprecedented increase in our computational power [94]. Computational methods in particular enable us to perform a large number of "experiments" *in silico*, with param-

eters varying in a wide range, in a reasonably short time. Such experiments allow us to explore parameter space more broadly and to make predictions about conditions that have not been tested before and, more importantly, are currently beyond the reach of our empirical power. Naturally, for such computational experiments to have any explanatory power, they must be validated thoroughly with behavioural data.

In this work, we use an agent-based model in which agents are controlled by artificial neural networks (ANNs) that differ in many important aspects from the more common ANN method. Because the logic of these networks is determined by logic gates with the Markov property we refer to these neural networks as Markov Brains [45]. Below, we describe the structure, function, and encoding of Markov Brains, but see [69] for a full description of their properties and how they are implemented. Markov Brains have been shown to be well-suited for modelling different types of behaviour observed in nature, from simple elements of cognition such as motion detection [184] and active categorical perception [116, 186], to swarming in predator-prey interactions [140], foraging [138], and decision-making strategies in humans [98].

5.4.1 Markov Brains

Markov Brains are networks of variables connected via probabilistic or deterministic logic gates with the Markov property. While we often term these variables "neurons", the state of the variable is more akin to a binary firing *rate*, that is, each neuron is a binary random variable (i.e., a bit) that may take two values: 0 for quiescent and 1 for firing. Fig. 5.8A shows a schematic of a simple Brain consisting of 12 neurons (labeled as 0-11) at two subsequent time points t and $t + 1$. The state of neurons in this example are updated via two logic gates. Fig. 5.8B shows a gate that takes inputs from neurons 0, 2, and 6 and writes the output into neurons 6 and 7. This logic gate produces output states of neurons 6 and 7 at time $t + 1$ given input states at time t. Each gate is defined by a probabilistic logic table in which the probability of each output pattern for a given input is specified. For example, in the probability table shown in Fig. 5.8C, p_{52} specifies the probability of obtaining output state $(N_6, N_7) = (1, 0)$ (a state with decimal representation '2') given input states

$(N_0, N_2, N_6) = (1, 0, 1)$ (decimal translation '5'), that is,

$$p_{52} = P(N_0, N_2, N_6 = 1, 0, 1 \rightarrow N_6, N_7 = 1, 0).$$

Since this gate takes 3 inputs, 2^3 possible inputs can occur, which are shown in eight rows. Similarly, this probabilistic table has four columns, one for each of the 2^2 possible outputs. The sum of the probabilities in each row must equal 1: $\sum_j p_{ij} = 1$. When using deterministic logic gates (such as in this study), all the conditional probabilities p_{ij} are zeros or ones. In general, Markov Brains can contain an arbitrary number of gates, with any possible connection patterns, and arbitrary probability values in logic tables [69]. As is clear from this example, we do not implement the update of the Brain state using probabilities that are conditional on the environmental state $\vec{E}_t$; rather, we update the joint state $(\vec{E}_t, \vec{S}_t)$.

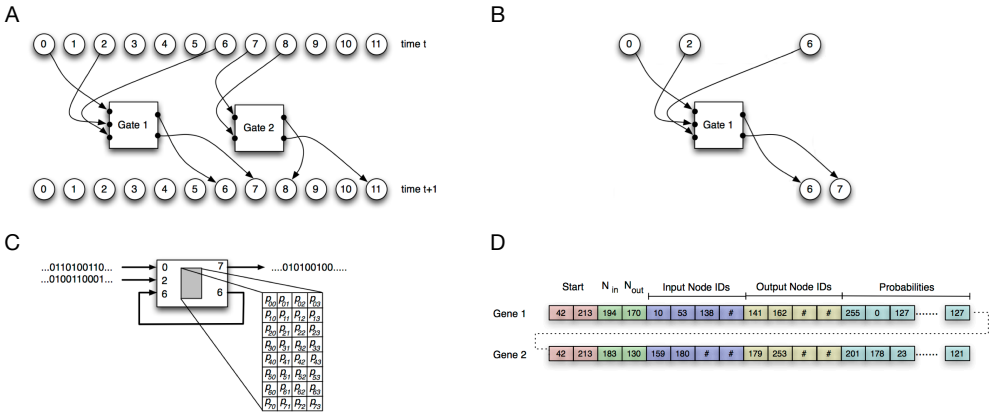

Figure 5.8: (A) A simple Markov Brain with 12 neurons and two logic gates at two consecutive time steps t and $t+1$. (B) Gate 1 of (A) with 3 input neurons and 2 output neurons. (C) Underlying probabilistic logic table of gate 1. (D) Markov Network Brains are encoded using sequences of numbers (bytes) that serve as agent's genome. This example shows two genes that specify the logic gates shown in (A), so that, for example, the byte value '194' that specifies the number of inputs N_{in} to gate 1 translates to '3' (the number of inputs for that gate).

In Markov Brains, a subset of the neurons is designated as sensory neurons that receive inputs from the environment. Similarly, another subset of neurons serves as actuator neurons (or decision neurons) that enable agents to take actions in their environment. In principle, an optimal Brain is designed in such a manner that a particular sequence of inputs (a time series of environmental states $\vec{\Sigma}_t = \vec{\sigma}_1, \vec{\sigma}_2, ..., \vec{\sigma}_t$) leads to a Brain state $\vec{S}_t$ that triggers the optimal response in that environment.

Rather than using an optimisation procedure that maximises an agent's performance over the probabilities $P(\vec{S}_t \rightarrow \vec{S}_{t+1}|\vec{E}_t)$, we use an evolutionary process in which a Brain's entire network is encoded in a genome [208] and optimisation occurs through the evolution of a population of such genomes using a "Genetic Algorithm" (GA, see for example [127]). In particular, each gene specifies a gate's connectivity and its underlying logic as shown in Fig. 5.8D. This evolutionary approach is explained in more detail in the following section.

5.4.2 Evolution of Markov Brains

Markov Brains can evolve to perform a variety of tasks representing different types of behaviours observed in nature. Selecting for any desirable task leads to the evolution of network connections and logic-gate properties that enable the agents to succeed in their environment. Each genome is a sequence of numbers ranging between 0-255 (bytes) that represent a set of genes that encode the logic and connectivity of the network. The arbitrary pair of bytes $\langle 42, 213 \rangle$ represents the "start codon" for each gate (Fig. 5.8D), while the downstream loci instruct the compiler how to construct the network, by encoding how many inputs and outputs define each logic gate, where the inputs come from (that is, which neuron or neurons), and where it writes to. In this manner, by "expressing" each gene, the network is fully determined via the connections between neurons and the logic those connections entail. Once a Brain is constructed, it is implanted in an agent whose performance is evaluated in an artificial environment that selects for the task. Those agents that perform best are rewarded with a differential fitness advantage. As these genomes are subject to mutation, heritability, and selection, they evolve in a purely Darwinian fashion (albeit asexually). The Genetic Algorithm specification details are shown in Table 5.2.

The population of Markov Brains evolves to judge the duration of an oddball tone ("longer" or "shorter") in multiple trials with different IOIs and oddball durations. The full set of all (IOI, standard tone), possible oddball tone durations, and the total number of trials for each pair of (IOI, standard tone) used in the evolution is shown in table 5.3. All told, there are 1,472 possible trials. However, agents are only evaluated on a subset of trials in every generation. This sampling

Table 5.2: Genetic Algorithm configuration. We evolved 50 populations of Markov Brains for 2,000 generations with point mutations, deletions, and insertions. We used roulette wheel selection, with 5% elitism, and with no cross-over or immigration.

Population size	100
Generations	2000
Initial genome length	5,000
Point mutation rate	0.5%
Gene deletion rate	2%
Gene duplication rate	5%
Elitism	5%

increases the evolution efficiency [24], and helps to avoid overfitting and enhances generalisation of learning [209]. In each generation, we randomly pick 22 trials from each (IOI, standard tone) pair (each row in Table 5.3) to form the evaluation subset: 11 trials with a longer oddball, and 11 trials with a shorter oddball, so as to prevent biasing Brains toward one response or the other. All agents of the population are then evaluated in that same subset of trials, which is 352 trials.

5.4.3 Experimental Setup

The Brains we evolve can have up to 16 neurons, of which one serves as the sensory neuron, and one delivers the decision (the "actuator" neuron). The remaining 14 neurons can be used for computation and signal transduction, but how many of them are actually used is determined by evolution. The population of Markov Brains evolves to judge the duration of a deviant tone (oddball) within a rhythmic sequence of otherwise identical tones, similar to experiments in [121] (see Fig. 5.9). In each trial, agents listen to a sequence of nine tones with a constant inter-onset-interval (IOI). An oddball is embedded within this sequence that is either shorter or longer in duration compared to the other eight tones (standard tones). Markov Brains sense the stimulus in one of their neurons (here, neuron 0, see Fig. 5.9). Agents must decide whether the oddball stimulus is longer or shorter than the standard tones. The agent is rewarded for correct duration judgements and does not gain any reward or incur a penalty for incorrect judgements. One neuron (neuron 15) in the Markov Brain is designated for delivering the decision ("longer" or "shorter").

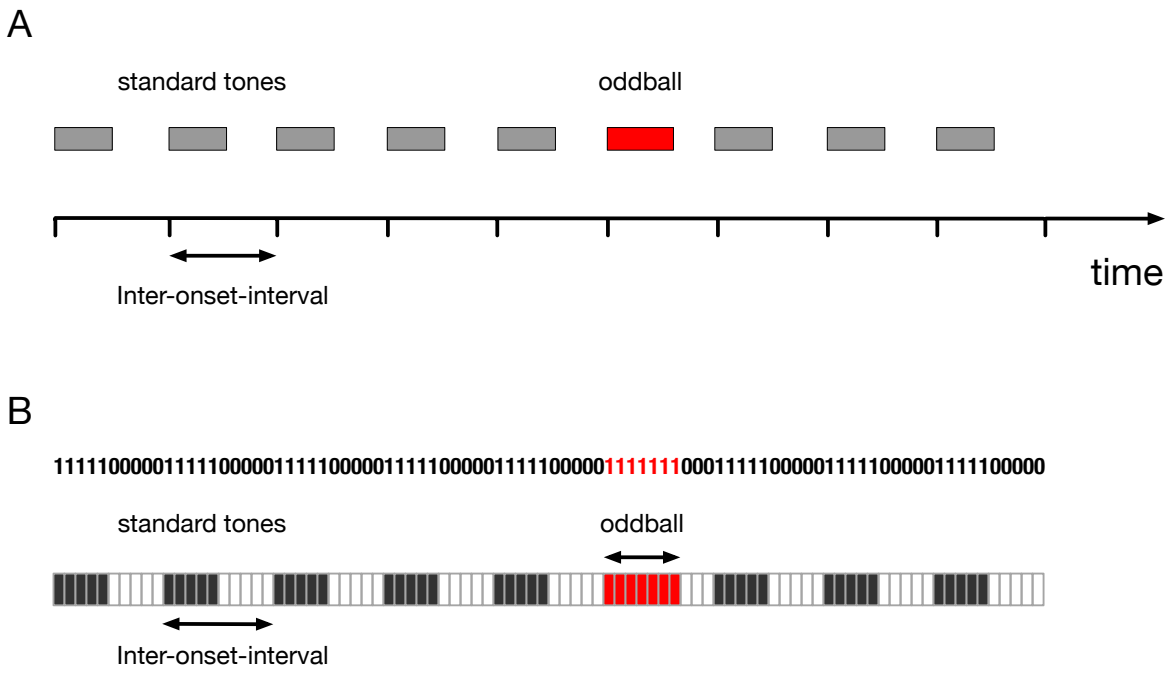

Figure 5.9: (A) A schematic of auditory oddball paradigm in which an oddball tone is placed within a rhythmic sequence of tones, i.e., standard tones. Standard tones are shown as grey blocks and the oddball tone is shown as a red block. (B) The oddball auditory paradigm, which is converted to a sequence of binary values, shown as sensed by the input neuron of a Markov Brain. When a stimulus is present, a sequence of '1's (shown by black blocks) is supplied to the sensory neuron while during silence, a sequence of '0' is fed to the sensory neuron. Each block shows one time step of the sequence experienced by the Brain.

For the purpose of fitness evaluation, agents are evaluated in several trials with different inter-onset-intervals (IOIs), different standard tones, a wide range of oddball durations, and with oddballs placed in different positions in the sequence. Standard tones range from 5 time steps to 12 time steps. The IOI is approximately twice the standard tone, and ranges from 10 to 25. Oddball durations can take any value from the shortest possible duration (1 time step) all the way to IOI minus 1 to avoid interfering with the next tone. During evolution, agents are not evaluated with oddball tones with the same duration as the standard tone since it is not shorter or longer than the standard tone. Oddballs can occur in either 5th, 6th, 7th, or 8th position, exactly as in the protocol of [121]. Our standard tones would be comparable in duration to those used in [121] if a digital time step is represented by a physical signal with about 70msec duration.

The set of all IOIs, standard tones, possible oddball-tone durations, and the total number of trials for each pair of (IOI, tone) is given in Table 5.3. All agents of the population are then evaluated in that same subset of trials, half of which with a longer oddball and the other half with shorter oddball, to avoid creating a bias in the agents' judgements. This subset of randomly picked trials consists of 512 trials (out of a total 2,852 trials): 22 trials for each (inter-onset-interval, standard tone) (see Table 5.3).

5.4.4 Discrete time in Markov Brains

The logic of Markov Brains is implemented by probabilistic or deterministic logic gates that update the Brain states from time t to time $t + 1$, which implies that time is discretised not only for Brain updates, but for the environment as well. Whether or not the brain perceives time discretely or continuously is a hotly debated topic [197], but for common visual tasks such as motion perception [198] discrete sampling of visual scenes can be assumed. For Markov Brains, the discreteness of time is a computational necessity. Because no other states (besides the neurons at time t) influence a Brain's state at time $t + 1$, the gates possess the Markov property (hence the name of the networks). Note that even though the Markov property is usually referred to as the "memoryless" property of stochastic systems, this does not imply that Markov Brains cannot have memory. Rather, memory can be explicitly implemented by gates whose outputs are written into the inputs of other gates, or even the same gates, i.e., to itself [45, 116].

5.4.5 Markov Brains as finite state machines

Because the Brains we evolve are deterministic, they effectively represent a deterministic finite-state automaton (DFA). There is considerable literature covering the mathematics of DFAs (see, for example [72]), but very little is applicable to the automata we evolve here. For example, realistic evolved automata are unlikely to have absorbing states, their stationary distributions are irrelevant, and they may be both cyclic and acyclic.

Table 5.3: Complete set of all inter-onset-intervals, standard tones, and oddball durations used for the evolution of duration judgement. Oddballs can occur in either of the 5th, 6th, 7th, or 8th position in the rhythmic sequence. Also, oddball durations are always either shorter or longer than the standard tone. The total number of trials for each pair $\langle \text{ioi}, \text{tone} \rangle$ is four times the IOI minus 2 (excluding oddball duration=standard tone, oddball duration=IOI), because the oddball can appear in four different positions within the rhythmic sequence.

(Inter-onset-interval, Standard tone)	Oddball tone durations	Total number of possible trials	Number of evaluation trials
(10, 5)	$\{1, 2, 3, 4\}$, $\{6, 7, 8, 9\}$	32	22
(11, 5)	$\{1, 2, \cdots, 4\}$, $\{6, \cdots, 10\}$	36	22
(12, 6)	$\{1, 2, \cdots, 5\}$, $\{7, \cdots, 11\}$	40	22
(13, 6)	$\{1, 2, \cdots, 5\}$, $\{7, \cdots, 12\}$	44	22
(14, 7)	$\{1, 2, \cdots, 6\}$, $\{8, \cdots, 13\}$	48	22
(15, 7)	$\{1, 2, \cdots, 6\}$, $\{8, \cdots, 14\}$	52	22
(16, 8)	$\{1, 2, \cdots, 7\}$, $\{9, \cdots, 15\}$	56	22
(17, 8)	$\{1, 2, \cdots, 7\}$, $\{9, \cdots, 16\}$	60	22
(18, 9)	$\{1, 2, \cdots, 8\}$, $\{10, \cdots, 17\}$	64	22
(19, 9)	$\{1, 2, \cdots, 8\}$, $\{10, \cdots, 18\}$	68	22
(20, 10)	$\{1, 2, \cdots, 9\}$, $\{11, \cdots, 19\}$	72	22
(21, 10)	$\{1, 2, \cdots, 9\}$, $\{11, \cdots, 20\}$	76	22
(22, 11)	$\{1, 2, \cdots, 10\}$, $\{12, \cdots, 21\}$	80	22
(23, 11)	$\{1, 2, \cdots, 10\}$, $\{12, \cdots, 22\}$	84	22
(24, 12)	$\{1, 2, \cdots, 11\}$, $\{13, \cdots, 23\}$	88	22
(25, 12)	$\{1, 2, \cdots, 11\}$, $\{13, \cdots, 24\}$	92	22

We define the state of a Markov Brain as the vector of states of all neurons except the sensory ones [166, 62, 159]: $\vec{S}_t = (N_p, N_{p+1}, ..., N_{n-1})$, where N_i is the state of the i^{th} neuron, p is the number of sensory (or peripheral) neurons, $(N_0, N_1, ..., N_{p-1})$ is the state vector of sensory neurons, and n is the total number of neurons. We abbreviate the Brain-state using the decimal translation of the state vector as:

$$S_t = \sum_{i=p}^{n-1} N_i(t) \times 2^i. \tag{5.2}$$

The Brain state can be thought of as a snapshot of the entire Brain that contains information about the activity (firing rate) of all neurons at that particular point in time. Markov Brains go through discrete states as the agent it controls behaves, reminiscent of what has been observed in monkeys performing a localisation task [166]. In our experimental setup, Markov Brains have 16 neurons in total, so $n = 15$. One of the neurons senses the stimulus, i.e. $p = 1$, so equation [5.2] can be written as $S_t = \sum_{i=1}^{15} N_i(t) \times 2^i$ which means the Brain can be in at most $2^{15} = 32,768$ different states. We also denote the sensory input at time t as $\vec{\sigma}_t$, and define the sequence of sensory inputs from time t_0 to t_1, $\vec{\Sigma}(t_0 : t_1) = (\vec{\sigma}_{t_0}, \vec{\sigma}_{t_0+1}, ..., \vec{\sigma}_{t_1})$.

The initial Brain state is always 0 since all neurons are quiescent at the outset. State-to-state transitions of an evolved Brain can be represented (or explained) as a mapping of the state of the Brain and the sensory input to the future state of the Brain. Formally, the set of all transitions of the Brain over all *visited* states in trials (states that Brains have taken on in those trials) can be viewed as a function $\mathcal{T}$ that takes the current state of the Brain S_t as well as the sensory input $\vec{\sigma}_t$ (in our experimental setup it is just one bit) as the input, and returns the future state of the Brain as the output, S_{t+1}:

$$\mathcal{T} : S_t, \vec{\sigma}_t \mapsto S_{t+1}, \quad \text{or} \quad S_{t+1} = \mathcal{T}(S_t, \vec{\sigma}_t), \tag{5.3}$$

We restrict the domain of variable S_t to those Brain states that actually occur during training (i.e., evolution) or test trials (early/late oddball tones). This function can be illustrated as a directed graph in which Brain states are represented by nodes (labelled by the decimal translation of the Brain state, see Eq. [5.2]) and edges represent transitions that are labelled with the stimulus that drives those transitions, $\vec{\sigma}$ (see [69] for a more detailed exposition of state-to-state diagrams).

5.4.6 Attention, experience, and perception in Markov Brains

We describe Markov Brains in terms of functions that take $(S_t, \vec{\sigma}_t)$ as the input and return S_{t+1} as the output.

Definition 1. If the Brain transitions from a particular state S_t to the same state S_{t+1} for all possible values of $\vec{\sigma}_t$ we say: the Brain *does not pay attention* to sensory input $\vec{\sigma}_t$ in state S_t.

Note that it is possible that the Brain does not pay attention to *parts* of the sensory input $\vec{\sigma}_t$ when the transition from S_t to S_{t+1} occurs independently of specific components of vector $\vec{\sigma}_t$. We emphasise that when the Brain does not pay attention to a sensory input in one transition, it does not imply that the stimulus is not sensed. Rather, it implies that even though sensed, the value does not affect the Brain's computation when in state S_t. It is crucial here that this definition of attention to a stimulus depends not only on the stimulus itself but also on the context in which it is sensed—this context is represented by the state S_t the Brain has reached. Because the Brain has reached the state S_t as a consequence of the temporal sequence of states traversed, this context is in fact historical. Also, note that the Brain state encompasses the actuator neuron (decision neuron), therefore, "not paying attention" is reflected in an agent's behaviour as well as the Brain's computations on sensory information. In a sense, the definition implies that an event that the Brain does not pay attention to should not alter its *experience* of the world, a concept that we will now define.

Definition 2. We define the Brain's experience of the environment (which is sensed as a sequence of sensory inputs $\vec{\Sigma}(0:t)$) as the sequence of Brain states it traverses, i.e., as $\vec{\chi}(0:t) = (\vec{S}_0, \vec{S}_1, \vec{S}_2, ..., \vec{S}_t)$.

This definition implies that the experiences of different individual Brains can be different when encountering the exact same sensory sequence, hence, experience is subjective [190, 189]. Furthermore, an agent may have experiences in which it does not take any actions on its environment (does not make any physical changes to itself or the world). Thus, dreaming or thinking are instances of such experiences in humans [172, 190, 189]. However, if the agent takes any actions in its environment, those actions become part of the experience by definition. For example, in our experimental setup Brains can only "take an action" in one particular time step of the trials. As a

result, a sequence of states that excludes that time step is still an experience, but does not involve any actions from the agent. It is also crucial to understand that the experience of the environment that is represented within Brain states is not just a naive projection of the world on the Brain, but rather contains integrated information about the relevant aspects of the environment (cues), while ignoring unimportant details (noise). In a very real sense, a Brain separates signal from noise; information from entropy [177].

In general, two different input sequences $\vec{\Sigma}_1(0:t)$ and $\vec{\Sigma}_2(0:t)$ will result in the Brain having two different experiences $\vec{\chi}_1(0:t)$ and $\vec{\chi}_2(0:t)$, but not necessarily. If experiences $\vec{\chi}_1(0:t)$ and $\vec{\chi}_2(0:t)$ are exactly the same, it means that (according to Definitions 1 and 2) the Brain does not pay attention to inputs during those transitions in which $\vec{\Sigma}_1$ and $\vec{\Sigma}_2$ are different. While in Definition 1 we only considered the Brain's transition at one time step, we can also look at the sequence of *future* Brain states, to discover how sensory inputs affect the Brain's computations and transitions multiple time steps after the input is sensed. Now, consider two input sequences $\vec{\Sigma}_1(0:t)$ and $\vec{\Sigma}_2(0:t)$ that differ in time steps $(0:t')$, where $t' < t$. Also, suppose $\vec{\Sigma}_1(0:t)$ and $\vec{\Sigma}_2(0:t)$ result in two different experiences $\vec{\chi}_1(0:t)$ and $\vec{\chi}_2(0:t)$. The effect of sub-sequence $\vec{\Sigma}(0:t')$ can be gauged by how different experiences $\vec{\chi}_1(0:t)$ and $\vec{\chi}_2(0:t)$ are as a result. For example, if two input sequences $\vec{\Sigma}_1(0:t')$ and $\vec{\Sigma}_2(0:t')$ (during time interval $0:t'$ where they are different) throw the Brain into two different regions in state space and therefore give rise to completely different experiences, then those inputs disturb experiences substantially. If, by contrast, $\vec{\Sigma}_1(0:t')$ and $\vec{\Sigma}_2(0:t')$ only result in different experiences temporarily (for example, during $0:t'$) while $\vec{\chi}_1$ and $\vec{\chi}_2$ become similar or identical later, then the differences in inputs is less disruptive to the Brain's experience. In particular, if the experiences have identical states at decision time t_d (assuming that $t_d \in [0:t]$), the differences in sensory inputs impact experiences $\vec{\chi}_1$ and $\vec{\chi}_2$ even less. We emphasise that the Brain state at the point of decision is key, because at this time point in the trial, the state of the Brain specifies the Brain's judgement, and more importantly, represents the path traversed in state space to reach this state. Consequently, we use the Brain state at decision time to define what it means to "perceive" a sensory input sequence.

Definition 3. If a Brain encounters two different input sequences $\vec{\Sigma}_1(0:t)$ and $\vec{\Sigma}_2(0:t)$, yet ends up in the *same* state S_t at decision time t in both cases, we say that the Brain *had the same perception* of sensory sequences $\vec{\Sigma}_1(0:t)$ and $\vec{\Sigma}_2(0:t)$.

By this definition, "having the same perception" is a superset of "having the exact same experience" when encountering two different sensory sequences. As discussed earlier, if the Brain has the exact same experience when exposed to two different input sequences, it clearly does not pay attention to the sub-sequence of the inputs that is not common between the two input sequences. In the same vein, how similar the experiences are for two different input sequences correlates with how little the Brain pays attention to those parts of input sequences that are not the same. This correlation captures the idea that there are different levels of "not paying attention" to a phenomenon in the environment. At the same time, it becomes clear that events that evoke the same perception (and thus similar experiences) must overlap in the significant parts of the sensory input. In this manner, the state of the Brain—being specific to the path in state space that leads to it—can encode "involuntary memory", in the same way as Marcel Proust's memories of the past [153] are triggered by the taste of a Madeleine dipped in Linden tea.

5.4.7 Information shared between perception and the oddball tone

Here we describe the procedures used to calculate the information shared between perception, (the Brain state at decision time-step), and the different oddball tone properties such as its duration, onset, and ending time-step. Markov Brains are tested against oddball tones varying in durations as well as different onsets with respect to the rhythm of the sequence. For each individual Brain we create an ensemble of trials with the same inter-onset-interval and standard tone, in which oddball tones differ in duration, onset, or both. We can calculate the information shared between the perception of each individual Brain and oddball properties for a given inter-onset-interval and standard tone using the standard Shannon information [38]

$$I(S_d : T_{ob}) = \sum_{s_d, t_{ob}} p(s_d, t_{ob}) \log\left(\frac{p(s_d, t_{ob})}{p(s_d)p(t_{ob})}\right), \tag{5.4}$$

where S_d denotes the Brain state at decision time (which we defined as perception) and T_{ob} denotes oddball properties, for example the oddball duration. The shared information between the perception and the oddball properties (duration, onset, and ending time-step) captures the correlation between the perception of the Brain and each of the oddball properties. It is noteworthy that perception occurs after the oddball tone has arrived and terminated. Thus, the information Eq. (5.4) measures how well each of the oddball tone properties can predict how the Brain perceives the tone.

5.5 Additional Experiments and Analysis

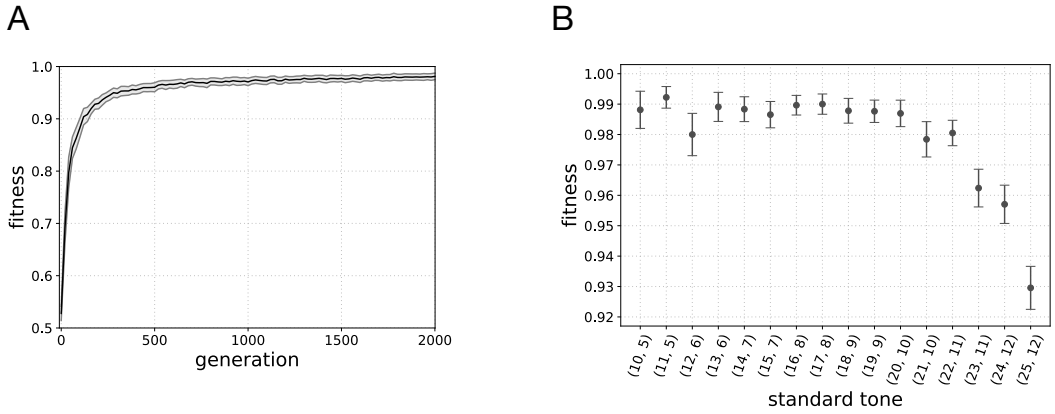

Figure 5.10: (A) Mean fitness across all 50 lineages and 95% confidence interval as a function of generation shown every 20 generations. (B) Mean fitness (and 95% intervals) of best agents picked from each of the 50 populations after 2000 generations as a function of inter-onset-interval, standard tone.

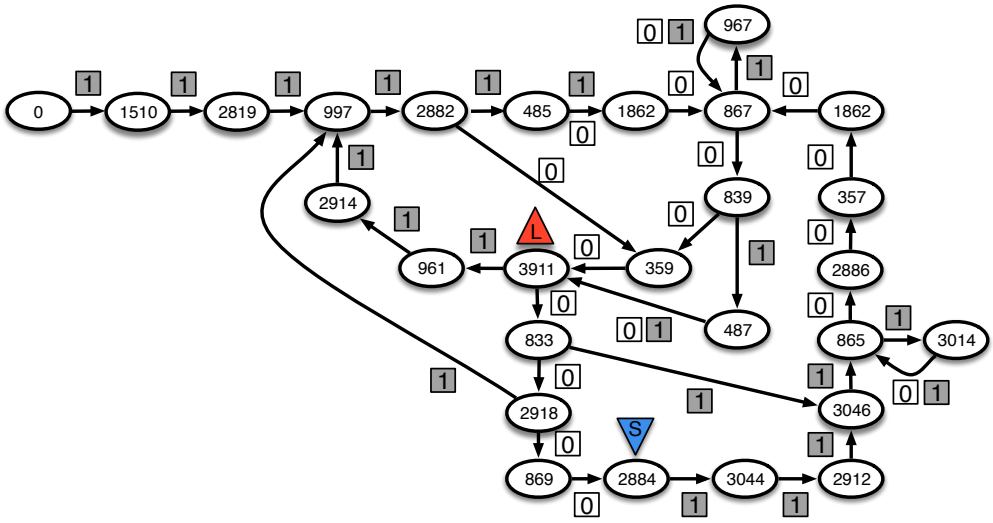

Figure 5.11: State-to-state transition diagram of a Markov Brain for IOI=10, standard tone=5, oddball tones=4 and 6, and onset of oddball tones can be 2 time steps early and 2 time step late.

5.5.1 Fitness landscape structure and historical contingencies result in Markov Brains using smaller regions of state space in trials with longer IOIs

In the main text we described that the judgement accuracy deteriorates as the IOI (and therefore tone lengths) increases. More specifically, even though the relative JND values remain in the same range for different IOI and standard tones (see Fig. 2B in the main text), PSE values start to deviate from the standard tone leading to higher values of "constant errors" (CE) that is, the difference between PSE and POE (see Fig. 2C in the main text). Here, we show that 1) deviations of PSEs in longer IOIs result from the fitness landscape structure and historical contingencies (see for example [59, 17]), and 2) the mechanistic basis of these deviations is associated with the size of the state-space Markov Brains use to encode stimuli characteristics.

As discussed before, Markov Brains display periodic firing patterns in response to rhythmic stimuli. These periodic patterns result in the formation of loops in their state transitions. This is the dominant mechanism by which Brains evolve to entrain to rhythmic stimuli, and encode temporal characteristics of the stimuli (i.e., rhythm and standard tone's duration). The distribution of the period of these periodic firing patterns, that is, the lengths of the loops in state transition diagrams is shown here again in Fig. 5.12A. Since the first four standard tones are provided so that Brains entrain to the rhythm, we measured the period of state transitions after the first four intervals, without an oddball tone. We also measured the number of distinct states each Brain visits during these periodic state transitions. Fig. 5.12B shows the distribution of number of distinct states in traversing loops during entrainment for 50 evolved Brains for each IOI. Note that these data represent number of distinct states in multiple loops, therefore, it is possible for a Brain to visit more states than the IOI. Note also that in traversing the loop once (in one period of the sequence) it is possible to visit some Brain states more than once. For example, the sequence: 6,3,1,1,6,3,1,1,... has a period of 4, but only three distinct states are visited. These results indicate that the number of distinct states visited by evolved Brains, i.e., the size of the state space used to encode temporal information, starts to plateau for longer IOIs.

The duration judgement task in trials with longer IOIs and standard tones is inherently more

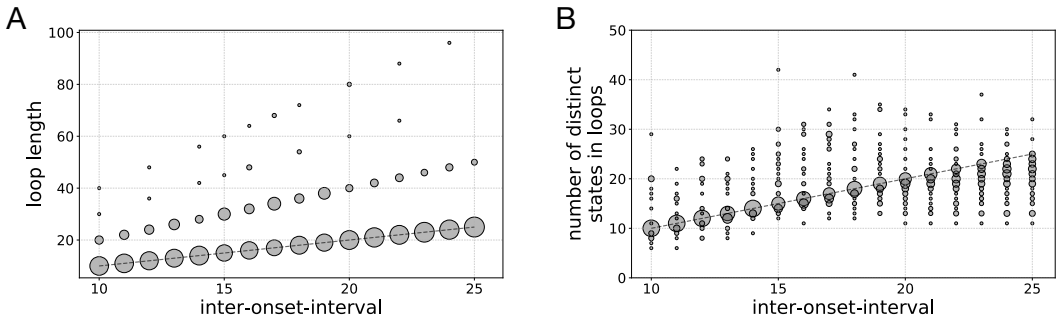

Figure 5.12: (A) The distribution of loop sizes of 50 evolved brain for each inter-onset-interval (IOI). The size of the markers is proportional to the number of Brains (out of 50) that evolve a particular loop lengths in each IOI. (B) The distribution of number of distinct states in loops visited by Markov Brains in a sequence of rhythmic standard tones, as a function of IOI. The dashed line shows the identity function line.

difficult (see Fig. 5.10B) for two reasons. First, longer rhythms and durations require more memory and computations to encode temporal information, and second, the number of possible oddball tones (in range $[1, IOI - 1]$) is greater in longer IOIs compared to the number of possible oddball tones in shorter IOIs. As a result, Markov Brains need to use progressively larger regions of their state-space to encode the temporal information and moreover, they need more evoltionary time to learn a larger number of patterns; however, state-space size does not grow linearly with IOI but rather begins to plateau (Fig. 5.12B) which, in turn, leads to less accurate performance in duration judgements in trials with longer rhythms and a systematic increase in PSE and CE values. This plateau in utilisation of state-space occurs not because of limitations in Markov Brains capacity but due to historical contingencies in the evolution. More specifically, the fitness landscape is structured in such a way that Markov Brains evolve to perform the duration judgement task for shorter IOIs earlier during the evolutionary course. As a consequence, algorithms that emerge later in evolution that perform the task in longer IOIs are built upon those algorithms evolved earlier. In order to provide further support for the claims we made here, we conducted a series of additional experiments. In the following sections we present results for the evolution of Markov Brains performing duration judgement for various experimental setups that differ slightly from the original experimental setup used in the main text.

5.5.1.1 Longer evolutionary time does not resolve systematic behavioural distortions in longer rhythms/standard tones

In the first set of additional experiment, we continued running the experiments presented in main text (which were run for 2,000 generations) for longer evolutionary time, namely 10,000 generations. Fig. 5.13 shows the fitness values of the best performing agents averaged across 50 runs as a function of IOI and colour-coded at different evolutionary times. We observe that the average fitness values increase in all IOI and standard tones with evolution, however, we still observe the same pattern that the performance drops as IOI increases. Fig. 5.14 shows CE values as a function of (IOI, standard tone) at different evolutionary time points. These results show that constant errors in longer IOIs decrease with evolutionary time, however, this decrease slows down considerably and more importantly, a similar trend in CE values vs. (IOI, standard tone) is observed in all generations.

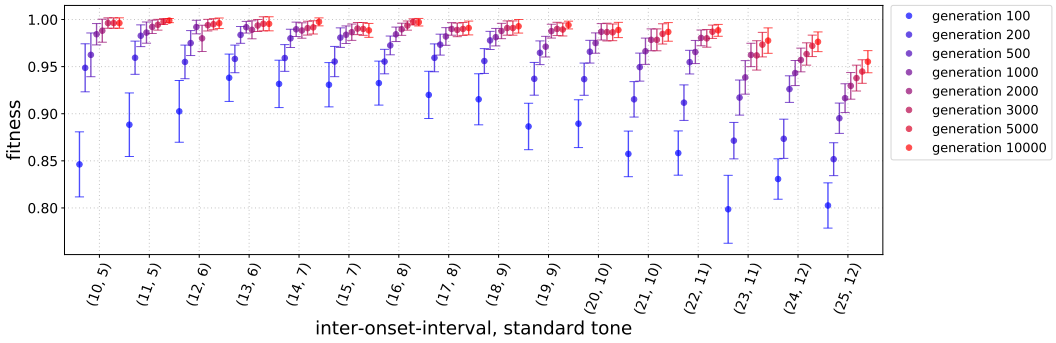

Figure 5.13: (A) Mean fitness across all 50 lineages and 95% confidence interval color-coded at different evolutionary times as a function of inter-onset-interval, standard tone.

Fig. 5.15 shows the number of distinct states used to encode temporal information corresponding to each IOI at different evolutionary time points. After 100 generations, the distributions of state-space size in shorter rhythms (IOIs 10-14) peak at the IOI (the identity function shown with dashed line) but as the IOI increases the peak of the distribution start to deviate from the identity line and begin to spread more widely. As evolution progresses, the distribution of distinct states in a larger number of IOIs peaks at the identity function but in all the plots shown in Fig. 5.15 (after different number of generations), the distributions that deviate from the identity line correspond to

the longest IOIs. For example, after 2,000 generations the distributions for IOIs 23-25 are further from the identity line, and after 10,000 elapsed generations this occurs for IOIs 24, and 25. Recall that we observed a similar pattern in CE values, where at the beginning of evolution CEs for shorter IOIs are around 0 but begin to deviate from 0 for longer IOIs, and as populations evolve further CEs for larger and larger number of IOIs approach 0. Note that the size of the state-space corresponding to each rhythm is indicative of how accurately the representation of that rhythm is encoded in the Brain. And clearly, in longer IOIs Markov Brains do not use as accurate an encoding and therefore, their performance drops for longer IOIs and CE values start to increase systematically.

Here we investigate in more depth the correlation between CE values and the size of state-space used by Markov Brains to encode temporal information. As discussed before, the optimum number of distinct states used to encode stimuli characteristics is the length of the rhythm, i.e., IOI. When the number of distinct states used to encode the rhythm length is smaller than IOI, it means that different time points during that interval have the same representation in the Brain because the Brain must visit some state(s) more than once (at different time points). For example, consider a Brain that is entrained to a rhythm and is traversing a loop in state-space. An oddball tone results in the Brain exiting that loop (we showed such an example in the main text). In this case, if the exit from the loop occurs from a repeated state in that loop, the Brain's experiences of oddballs that end at different time points would be exactly the same. Alternatively, when the number of distinct states visited when traversing the loops is greater than IOI, it means that the period of that loop is not IOI but a multiple of the IOI. This may also result in less accurate performance in duration judgement task, for example in the judgement of oddballs with the same duration that occur in different positions (recall that oddball tones can occur at 5^{th}, 6^{th}, 7^{th}, or 8^{th} position).

In Fig. 5.12B, we observed that the distribution of number of distinct states in loops peaks at IOI for shorter IOIs at the outset of evolution and increasingly more distributions move towards the IOI and accumulate around IOI. Let $\vec{D}_{\text{IOI}} = (d^1_{\text{IOI}}, d^2_{\text{IOI}}, d^3_{\text{IOI}}, \ldots, d^N_{\text{IOI}})$, where d^i_{IOI} represents the number of distinct states the i^{th} Brain uses in its loops for a particular IOI, and $N = 50$ since we have 50 evolved Brains. Thus, each distribution in Fig 5.15 can be represented by a vector $\vec{D}$. We

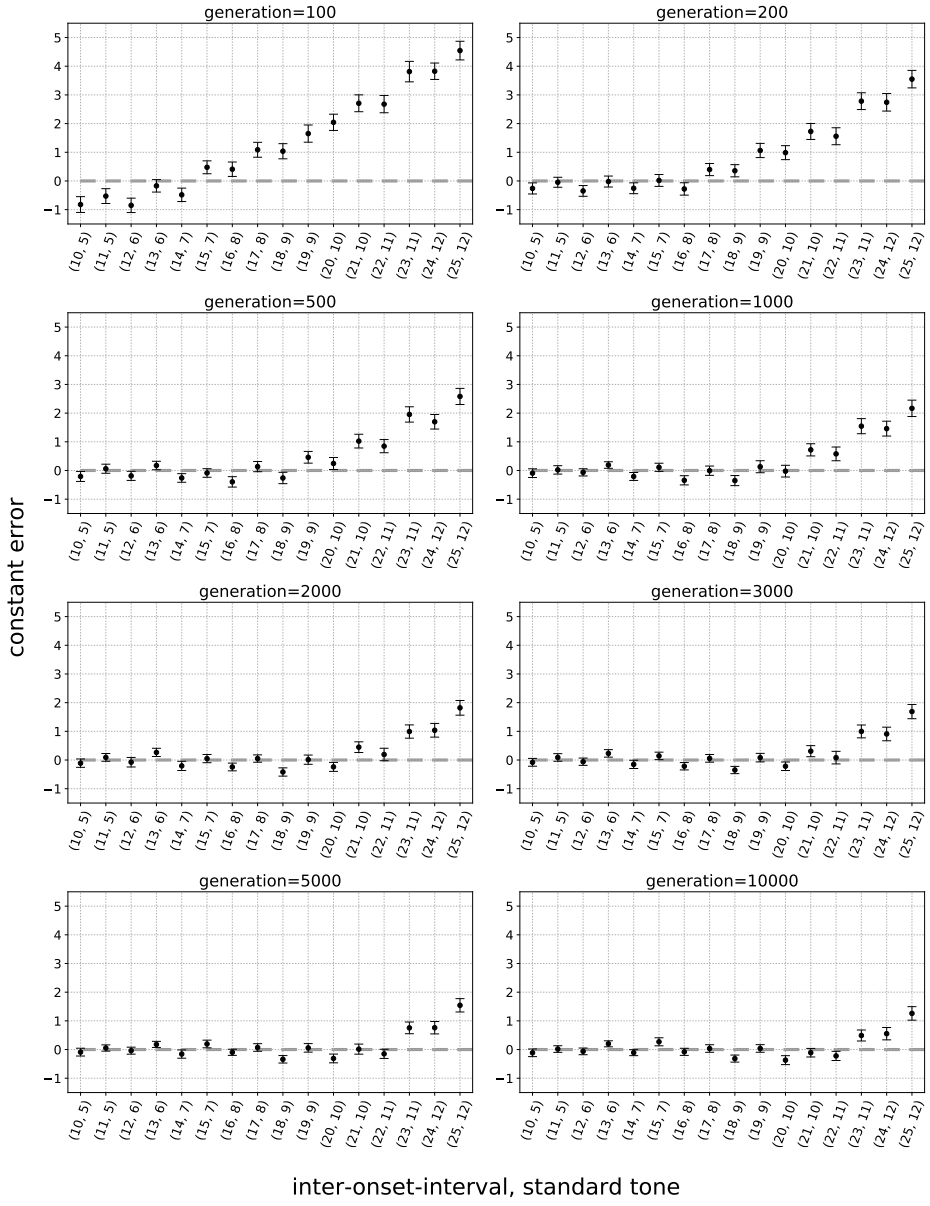

Figure 5.14: Constant errors and their 95% confidence interval for 50 best performing Brains as a function of inter-onset-interval, standard tone at different evolutionary times. Dashed line shows zero constant error.

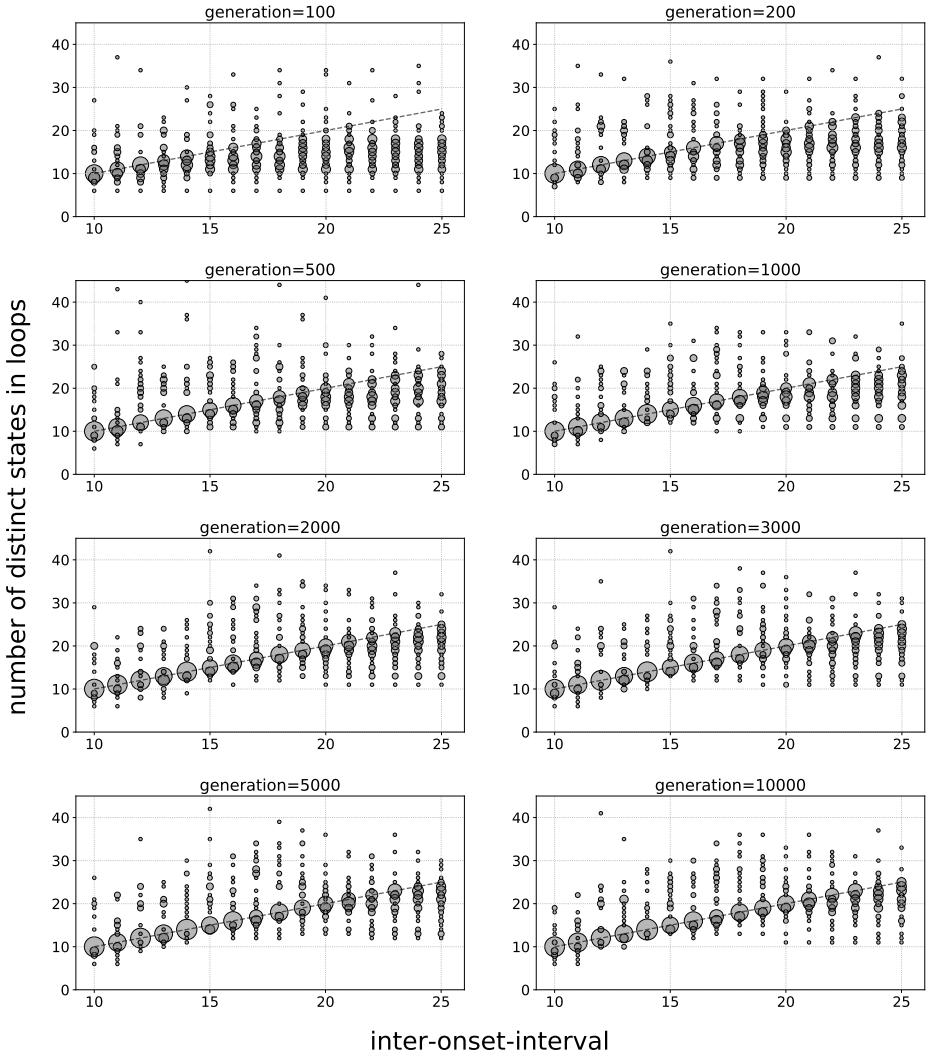

Figure 5.15: The distribution of number of distinct states used to encode rhythm and standard tone duration, i.e., the number of distinct states in each loop, as a function of inter-onset-interval at different evolutionary times. The size of the circle is proportional to the likelihood at that loop size.The dashed line shows the identity function.

now calculate the distance of each distribution to the IOI by:

$$\delta_{\mathrm{IOI}} = \|d_i - \mathrm{IOI}\|_0 = \lim_{p \to 0} \left(\sum_i (d_i - \mathrm{IOI})^p \right)^{\frac{1}{p}}, \tag{5.5}$$

in which $\|\|_0$ denotes the ℓ_0-norm of vector $(d^1_{\mathrm{IOI}} - \mathrm{IOI}, d^2_{\mathrm{IOI}} - \mathrm{IOI}, \ldots, d^N_{\mathrm{IOI}} - \mathrm{IOI})$. In fact, δ_{IOI} simply reflects how many of the 50 Brains do not use exactly IOI distinct states in their loops. We calculated δ_{IOI} for each IOI and at different points in evolutionary time. We then normalised these δ_{IOI} by the maximum δ_{IOI} value. Fig 5.16 shows absolute CE values as a function of normalised δ_{IOI}. Each data point shown in grey represents δ_{IOI} calculated in a distribution at a specific evolutionary time and a particular IOI in Fig. 5.15).

We used a non-linear regression analysis [15] to find the correlation between the CE and δ_{IOI}. Since a large number of data points fall around CE=0 and in the lower range of δ_{IOI} (which is not surprising since most trials result in CEs that are not significantly different from 0), we applied binning with constant bin size to this data. Mean values of binned data and their standard deviations as well as the fitted function are also shown in Fig. 5.16. We tested three different kernel functions for regression analysis: 1) quadratic function, 2) ramp function, 3) softplus function ($f(x) = log(1 + e^x)$, which is a differentiable approximation of ramp function). Table 5.4 shows the regression analysis results for three different kernel functions. We compare these three models using Bayesian information criterion (BIC) [155]. These results show that the softplus function describes the pattern in the data better than quadratic and ramp function. This pattern can be interpreted as: there is no significant change in CE values for a range of small δ_{IOI}s, however, by further increasing δ_{IOI}, at some threshold CEs start to increase linearly with δ_{IOI}.

5.5.1.2 Training Markov Brains equally in all IOIs and standard tones has a minor effect on behavioural deviations in longer rhythms

In this experimental setup, we used the same set of inter-onset-intervals, standard tones, and oddball tones as used in original experimental setup. The only difference is that the number of evaluations for each (IOI, standard tone) is not constant anymore (in the original setup we evaluate Brains in 22

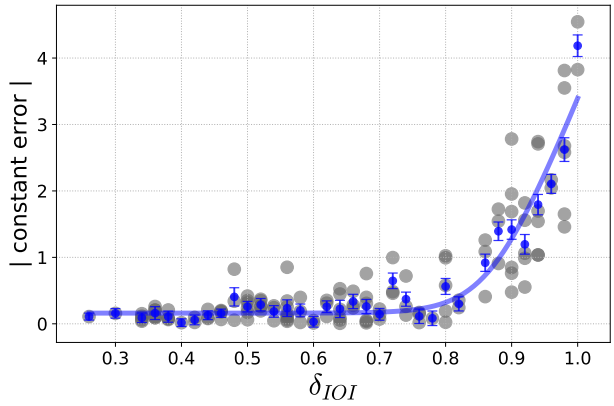

Figure 5.16: Absolute constant errors (CE) shown in grey as a function of δ_{IOI}, as well as the binned data and the fitted softplus curve.

Table 5.4: Non-linear regression analysis used to explain the correlation between the constant errors (CE) and δ_{IOI} which is a function of the distinct number of states used in encoding stimuli. Residuals sum of squares (RSS), and the Bayesian information criterion. A BIC difference > 10 provide very strong support for one model over the other [155].

function	RSS	BIC	ΔBIC with quadratic	ΔBIC with ramp	ΔBIC with softplus
quadratic	6.49	-48.29	0	-	-
ramp	2.41	-83.02	34.73	0	-
softplus	1.9	-91.39	43.10	8.37	0

trials for each IOI, standard tone) but in this modified setup it increases with IOI linearly. Table 5.5 shows the number of evaluation trials as well as IOI, standard tone, and total number of trials for each (IOI, standard tone). Note that we tried to keep the total number of evaluations in this setup, 368 (37.1% of all possible trials), as close as possible to that of the original setup 352 (35.5% of all possible trials). Note also that the number of evaluations in each (IOI, standard tone) is chosen proportionate to the number of oddball tones in that (IOI, standard tone).

Fig. 5.17 shows CE values for this experimental setup as a function of (IOI, standard tone) at different evolutionary time points in the experiments. It is evident that the same trend in CE values that was observed in the original setup can be seen in these experiments too. In particular, after 2,000 generations CEs for (IOI, standard tone)={(23, 11), (24,12), (25,12)} are significantly

Table 5.5: Complete set of all inter-onset-intervals, standard tones, and oddball durations used for evolution of duration judgement task. Oddballs can occur in either of 5th, 6th, 7th, or 8th position in the rhythmic sequence. Also, oddball durations are always either shorter or longer than the standard tone.

Inter-onset-interval	Standard tone duration	Oddball tone duration	Total number of possible trials	number of evaluation trials
10	5	{1, 2, 3, 4} , {6, 7, 8, 9}	32	8
11	5	{1, 2, $\cdots$, 4} , {6, $\cdots$, 10}	36	10
12	6	{1, 2, $\cdots$, 5} , {7, $\cdots$, 11}	40	12
13	6	{1, 2, $\cdots$, 5} , {7, $\cdots$, 12}	44	14
14	7	{1, 2, $\cdots$, 6} , {8, $\cdots$, 13}	48	16
15	7	{1, 2, $\cdots$, 6} , {8, $\cdots$, 14}	52	18
16	8	{1, 2, $\cdots$, 7} , {9, $\cdots$, 15}	56	20
17	8	{1, 2, $\cdots$, 7} , {9, $\cdots$, 16}	60	22
18	9	{1, 2, $\cdots$, 8} , {10, $\cdots$, 17}	64	24
19	9	{1, 2, $\cdots$, 8} , {10, $\cdots$, 18}	68	26
20	10	{1, 2, $\cdots$, 9} , {11, $\cdots$, 19}	72	28
21	10	{1, 2, $\cdots$, 9} , {11, $\cdots$, 20}	76	30
22	11	{1, 2, $\cdots$, 10} , {12, $\cdots$, 21}	80	32
23	11	{1, 2, $\cdots$, 10} , {12, $\cdots$, 22}	84	34
24	12	{1, 2, $\cdots$, 11} , {13, $\cdots$, 23}	88	36
25	12	{1, 2, $\cdots$, 11} , {13, $\cdots$, 24}	92	38

Table 5.6: Non-linear regression analysis used to explain the correlation between the constant errors (CE) and δ_{IOI} which is a function of the distinct number of states used in encoding stimuli. Residuals sum of squares (RSS), and the Bayesian information criterion.

function	RSS	BIC	ΔBIC with quadratic	ΔBIC with ramp	ΔBIC with softplus
quadratic	5.36	-66.42	0	-	-
ramp	1.59	-113.77	47.35	0	-
softplus	1.4	-118.65	52.23	4.88	0

different from 0 and similarly, after 10,000 generations the CE for (25, 12) is significantly different from 0. Fig. 5.18 shows state-space sizes as a function of IOI at different evolutionary time points. Similar to trends observed in the original setup, state-space sizes plateau as IOIs increase and again, their distributions are slightly closer to the identity function (dashed line) but not significantly so. Thus, we conclude that having the same training set size for all IOIs has little to do with distorted behaviours in longer rhythms. Fig. 5.19 shows the binned CE values as a function of δ_{IOI} as well as the fitted softplus function. We performed the non-linear regression analysis described before for this experiment and the results are presented in Table 5.6. Similar to previous experiment, the softplus function describes the pattern in CE values and δ_{IOI} better than the other two models.

5.5.1.3 Constant errors in longest rhythms are greater than zero regardless of trial size

In order to show that the deviations of PSE (from the point of objective equality, i.e., standard tone) in longer IOI, and standard tones is not specific to a particular value of IOI or standard tone, we used two experimental setups where one has a smaller set of (IOI, standard tone) with shorter IOIs and standard tone durations, and one that has a larger set of (IOI, standard tone) with longer rhythms, standard tones. The first training set is similar to the original experimental setup but we excluded trials with the following inter-onset-intervals and standard tones from the original setup: $\{(23, 11), (24, 12), (25, 12)\}$. Similar to the original setup, oddball tones can vary from 1 to IOI-1. In this experimental setup, there are 728 possible trials and all agents are evaluated in 20 trials from each IOI and standard tone (10 with longer and 10 with shorter oddball tones) which is 35.7% of all possible trials (in the original setup evaluation trials set was 35.5% of all possible trials).

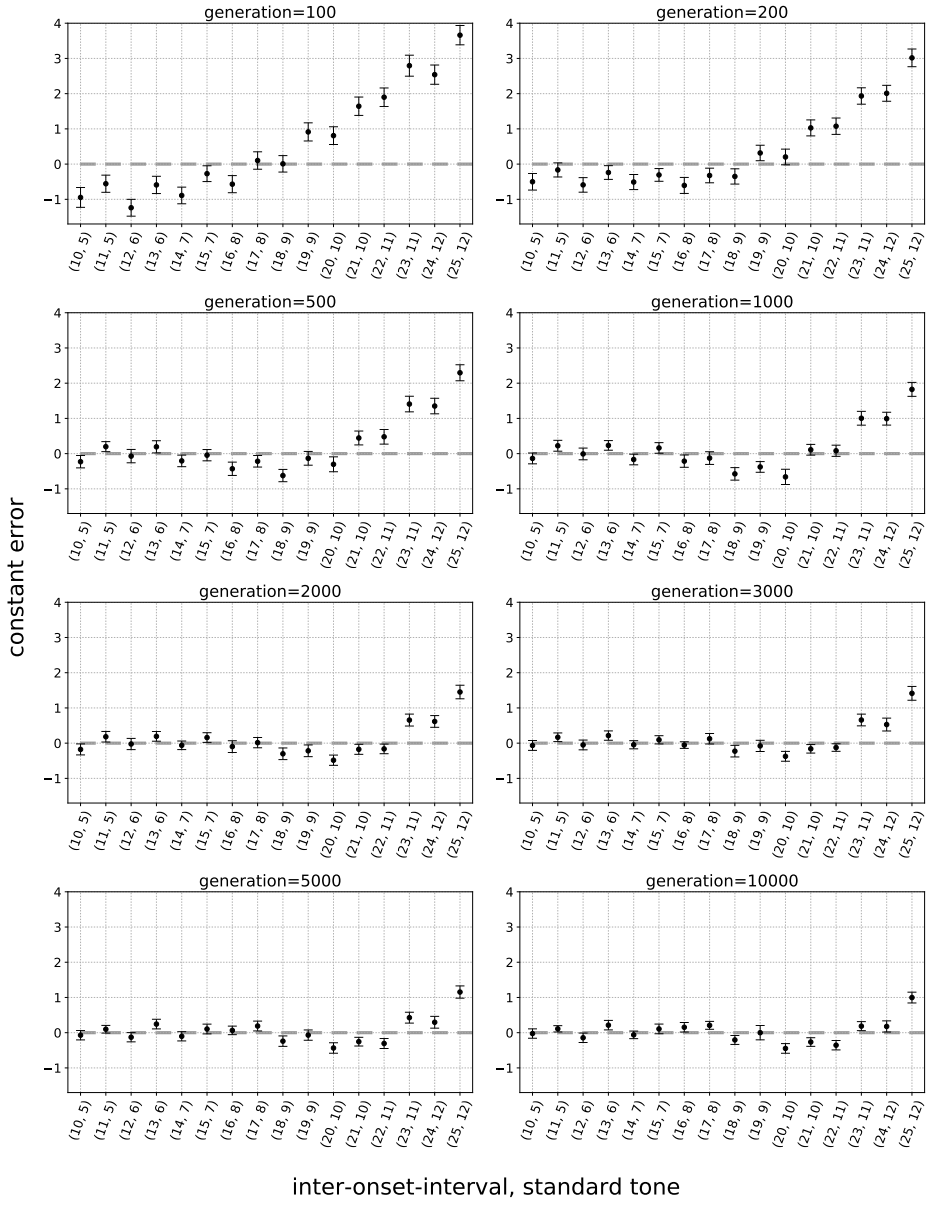

Figure 5.17: Constant errors and their 95% confidence interval for 50 best performing Brains as a function of inter-onset-interval, standard tone at different evolutionary times. Dashed line shows zero constant error.

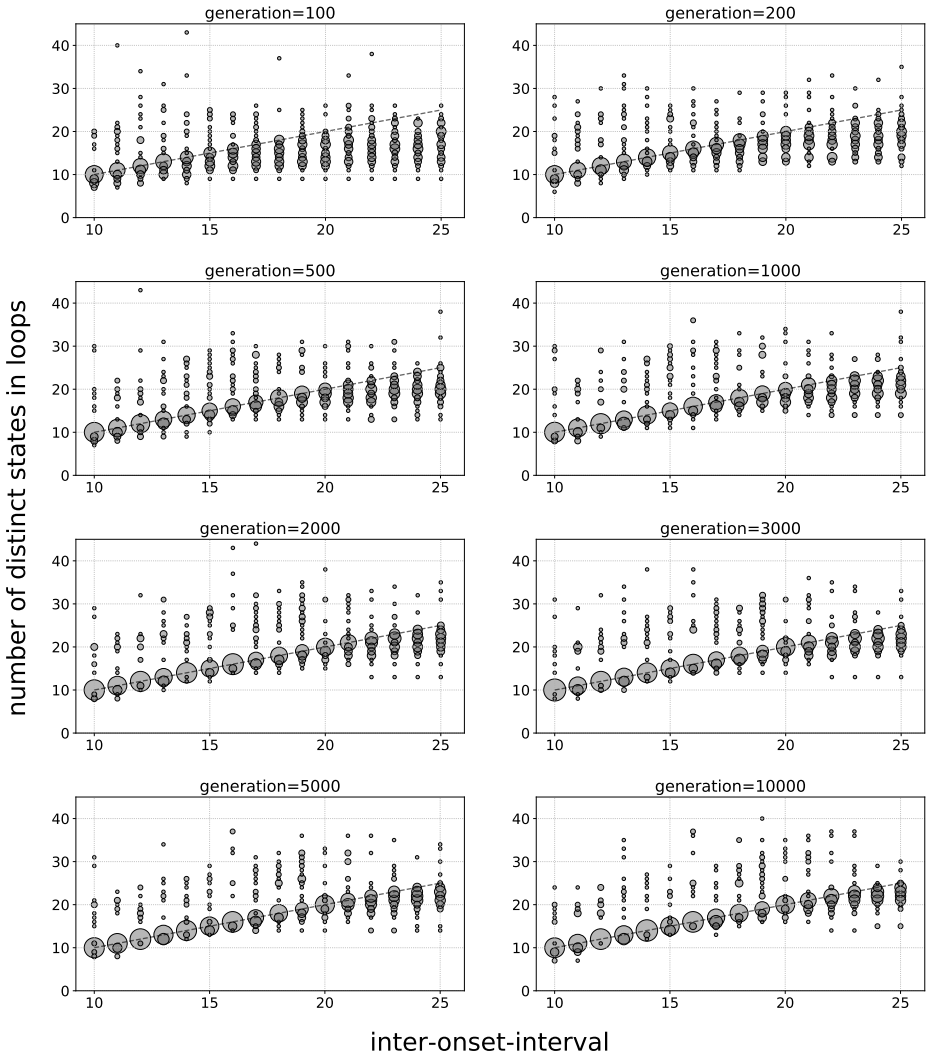

Figure 5.18: The distribution of number of distinct states used to encode rhythm and standard tone duration, i.e., the number of distinct states in each loop, as a function of inter-onset-interval at different evolutionary times. The dashed line shows the identity function.

Fig. 5.20 shows mean constant errors as a function of standard tones at different evolutionary times for this experimental setup. The increase in CEs is again observed for longer IOIs and noticeably, after 2000 generations in trials with (IOI, standard tone)={(10, 5), (11,5)}, all 50 Brains perform the duration judgement task perfectly (100% performance for all oddball tones in those rhythms) and we observe Brains perform the duration judgement task perfectly in more IOIs, and standard tone in later generations, for example after 10,000 generations Brains perform perfectly in (IOI, standard tone)={(10, 5), (11, 5), (12, 6), (14,7)}. Cognitive scientists and psychophysicists

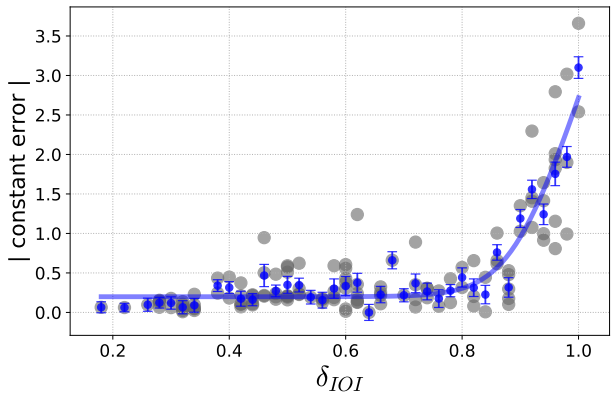

Figure 5.19: Absolute constant errors (CE) shown in grey as a function of δ_{IOI}, as well as the binned data and the fitted softplus curve.

are not in general interested in "trivial" experiments in which all the subjects answer 100% of questions correctly; therefore, we did not design our experimental setup such that Brains evolve to achieve 100% fitness either. Fig. 5.21 shows state-space size distributions as a function of IOI for different evolutionary time points. It is again evident that the state-space sizes start to plateau for longer IOIs but of course, not as drastically as in the original setup. The CE values, as well as binned means and their standard deviations, are shown as a function of δ_{IOI} are shown in Fig. 5.22. In Fig. 5.22, the blue dashed line shows the fitted softplus function. The results of the non-linear regression analysis are shown in Table 5.7. We again observe that the softplus function describes the pattern in CE values and δ_{IOI} better than the other two functions.

The second experimental setup has all the trials from the original and we also added the following inter-onset-intervals and standard tones: $\{(26, 13), (27, 13), (28, 14), (29, 14)\}$. In this experimental setup, there are 1400 possible trials and all agents are evaluated in 24 trials from each IOI and standard tone (12 with longer and 12 with shorter oddball tones) which is 34.3% of all possible trials to maintain the same ratio of evaluation trials to all possible trials. Fig. 5.23 shows mean constant errors as a function of standard tones at different evolutionary times for this experimental setup. These results show a similar pattern in CE values and more importantly, we observe that the CEs for the inter-onset-interval and standard tones $\{(23, 11), (24, 12), (25, 12)\}$ are not significantly different from 0 whereas in the original experiment, CEs were significantly

Table 5.7: Non-linear regression analysis used to explain the correlation between the constant errors (CE) and δ_{IOI} which is a function of the distinct number of states used in encoding stimuli. Residuals sum of squares (RSS), and the Bayesian information criterion.

function	RSS	BIC	ΔBIC with quadratic	ΔBIC with ramp	ΔBIC with softplus
quadratic	2.91	-76.33	0	-	-
ramp	1.48	-99.97	23.64	0	-
softplus	0.98	-114.51	38.18	14.54	0

Table 5.8: Non-linear regression analysis used to explain the correlation between the constant errors (CE) and δ_{IOI} which is a function of the distinct number of states used in encoding stimuli. Residuals sum of squares (RSS), and the Bayesian information criterion.

function	RSS	BIC	ΔBIC with quadratic	ΔBIC with ramp	ΔBIC with softplus
quadratic	7.03	-53.20	0	-	-
ramp	1.03	-126.34	73.14	0	-
softplus	0.89	-131.92	78.72	5.58	0

different from 0 in the same trials, i.e., $\{(23, 11), (24, 12), (25, 12)\}$. Fig. 5.24 shows state-space size distributions as a function of inter-onset-intervals for different evolutionary time points. We again observe that the state-space sizes start to plateau for longer IOIs but of course, but not as drastically as in the original setup. We performed the non-linear regression analysis on these data as well and the results are shown in Table 5.8. As observed in previous results, the softplus function describes the pattern in CE values and δ_{IOI} better than the other two models. The CE values, the binned data mean and standard deviations, and the fitted softplus function is shown in Fig. 5.25.

These results reaffirm that the entrainment and duration judgement task become much more difficult for longer (IOIs, standard tone) and with greater set of trials, and that furthermore, Markov Brains do have the capacity to use greater regions of the state-space and perform more accurately in longer IOIs. However, the historical contingencies in such fitness landscapes lead to less accurate strategies in duration judgements in longer IOIs which results from using smaller regions in state-space.

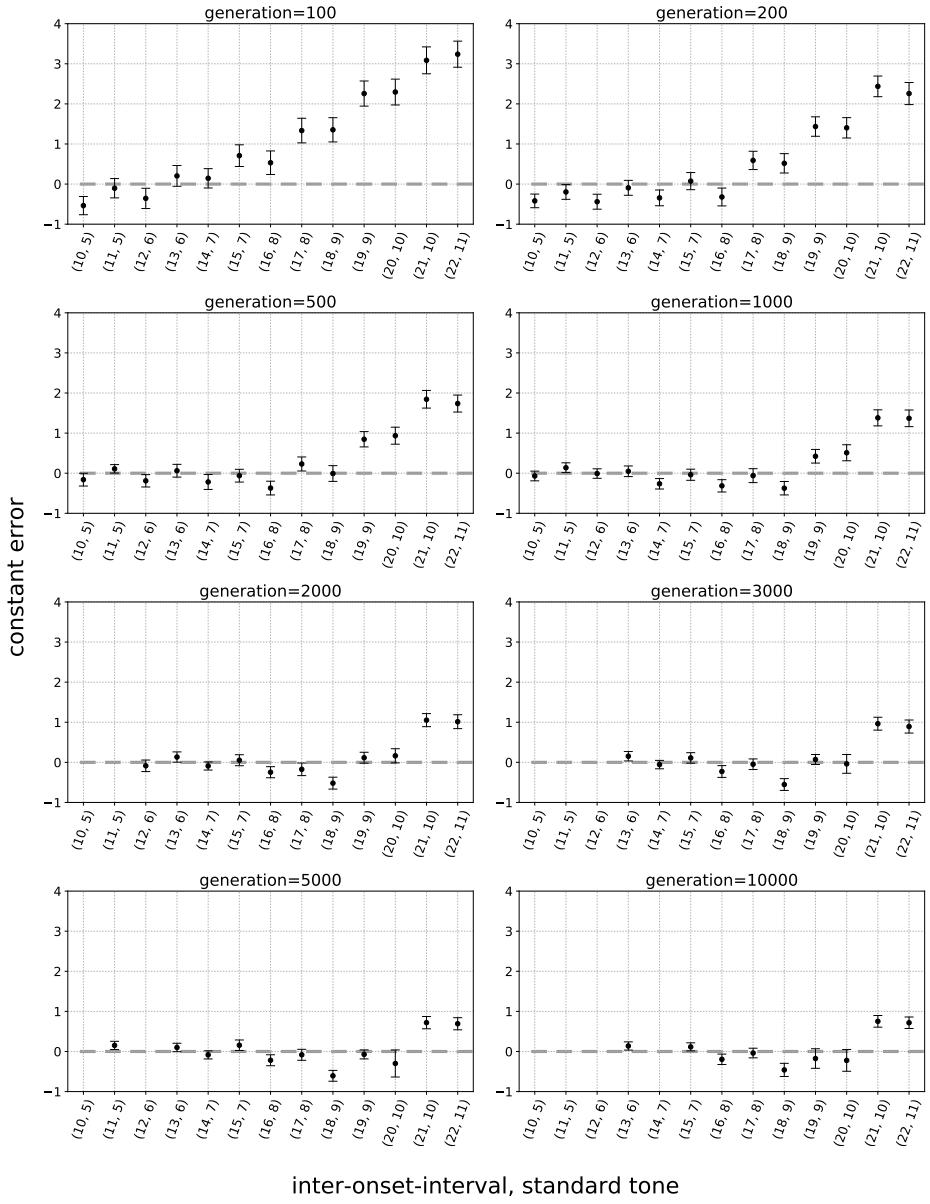

Figure 5.20: Constant errors and their 95% confidence interval for 50 best performing Brains as a function of inter-onset-interval, standard tone at different evolutionary times. There are some missing data points in these plots which is due to the fact that in those trials the performances of all 50 Brains are 100%, as a result, PSE would be exactly equal to the standard tone and the slope of the psychometric function would be infinity. Dashed line shows zero constant error.

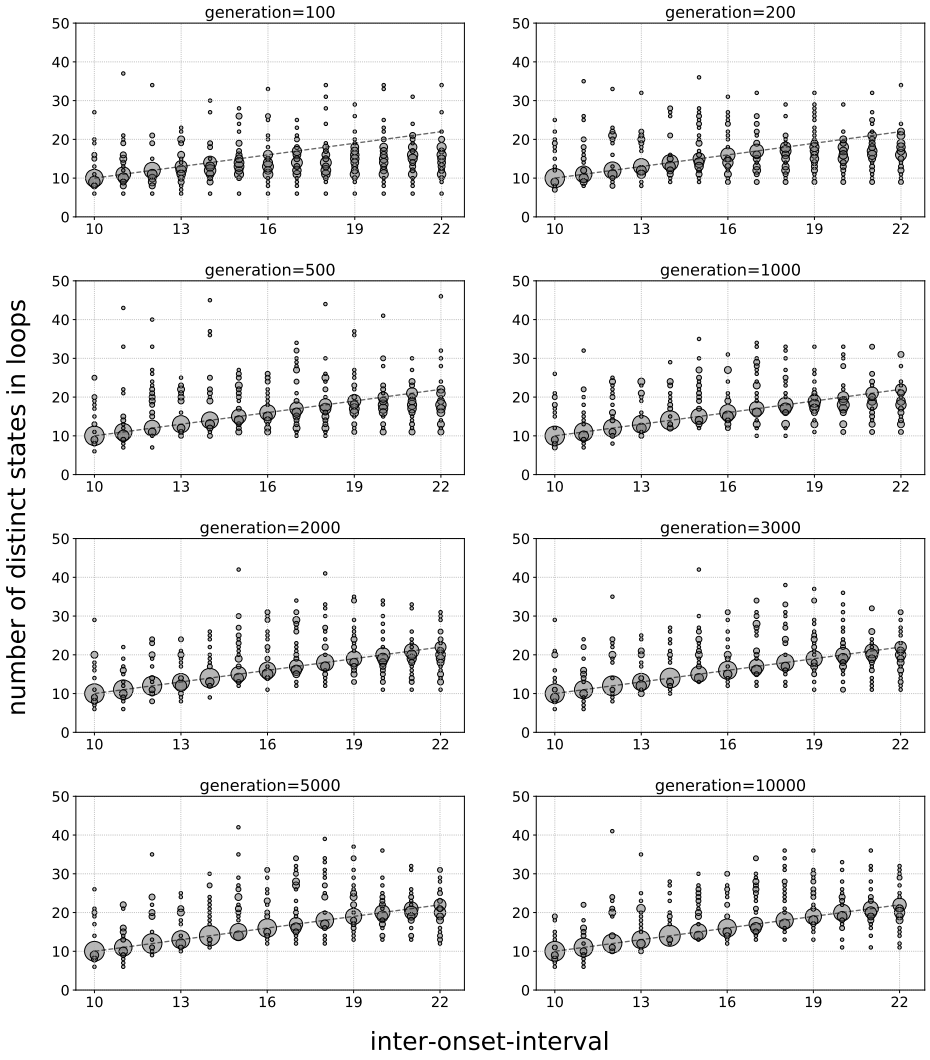

Figure 5.21: The distribution of number of distinct states used to encode rhythm and standard tone duration, i.e., the number of distinct states in each loop, as a function of inter-onset-interval at different evolutionary times. The dashed line shows the identity function.

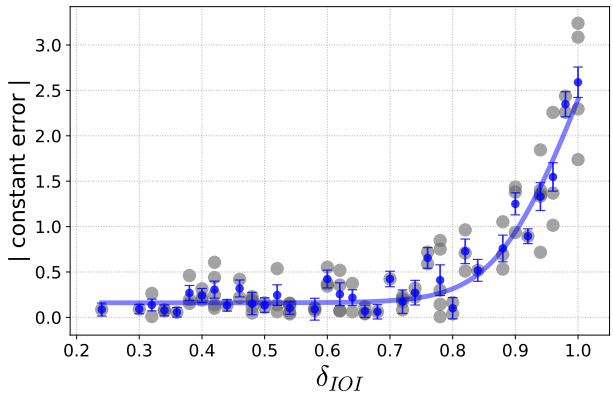

Figure 5.22: Absolute constant errors (CE) shown in grey as a function of δ_{IOI}, as well as the binned data and the fitted softplus curve.

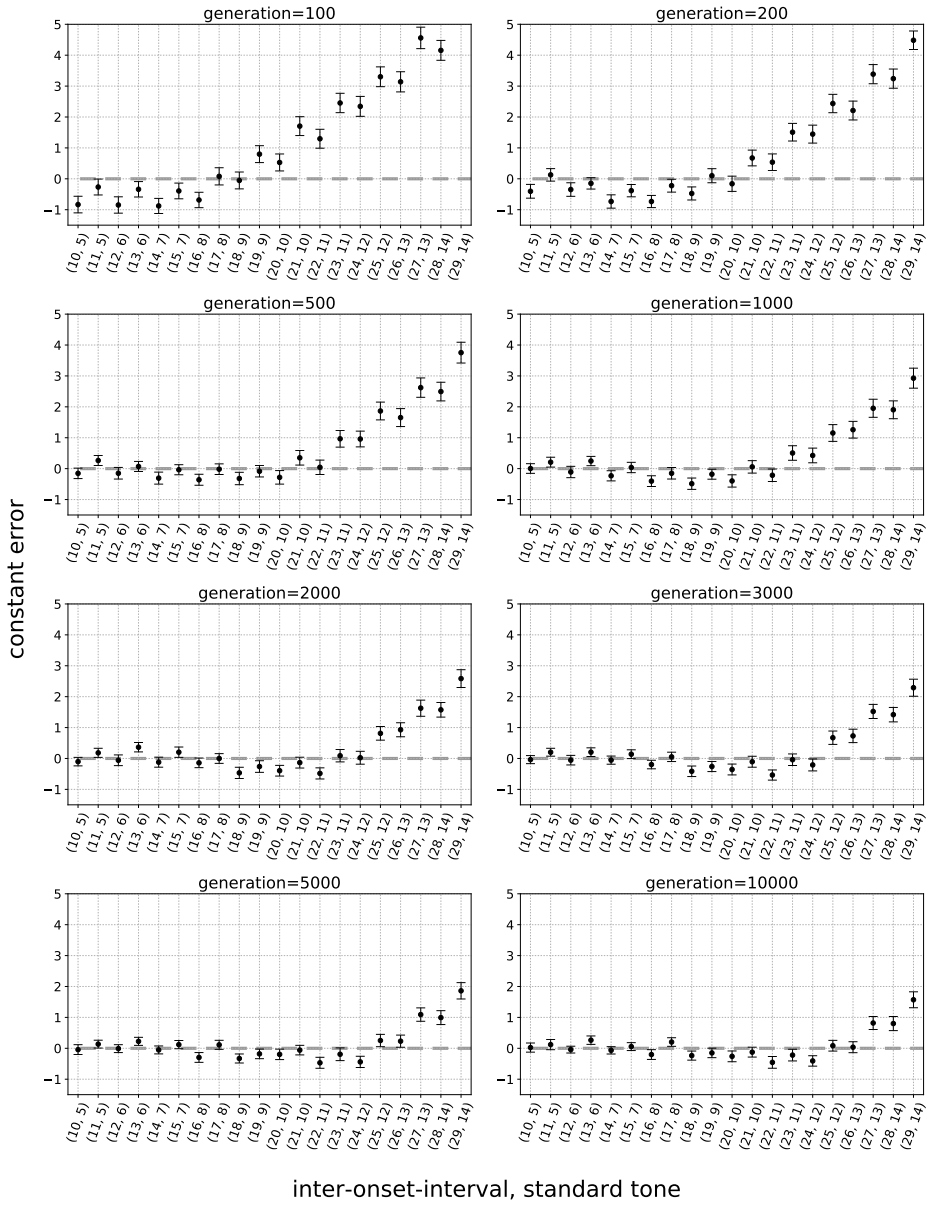

Figure 5.23: Constant errors and their 95% confidence interval for 50 best performing Brains as a function of inter-onset-interval, standard tone at different evolutionary times. There are some missing data points in these plots which is due to the fact that in those trials the performances of all 50 Brains are 100%, as a result, PSE would be exactly equal to the standard tone and the slope of the psychometric function would be infinity. Dashed line shows zero constant error.

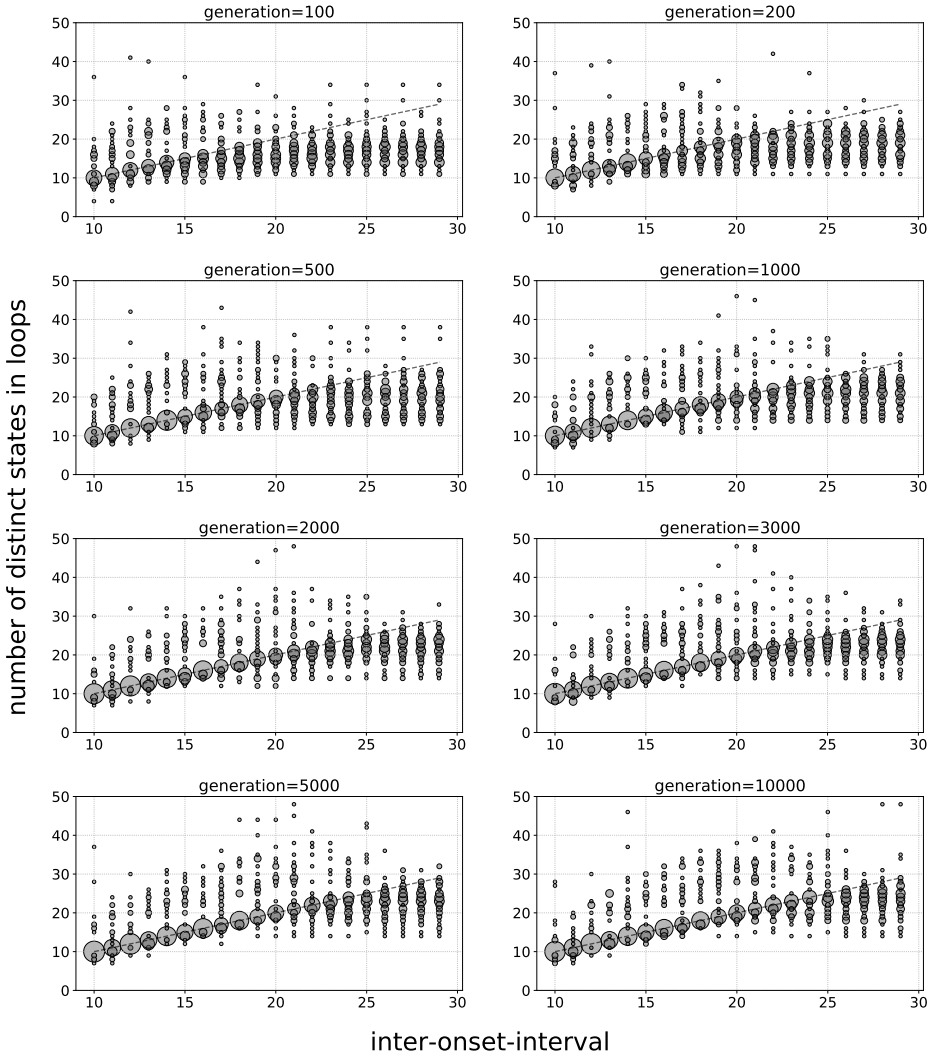

Figure 5.24: The distribution of number of distinct states used to encode rhythm and standard tone duration, i.e., the number of distinct states in each loop, as a function of inter-onset-interval at different evolutionary times. The dashed line shows the identity function.

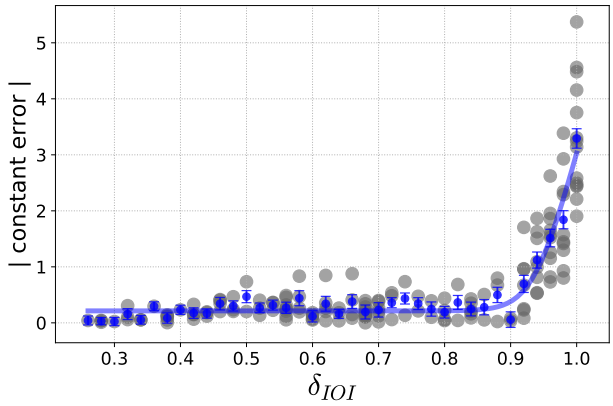

Figure 5.25: Absolute constant errors (CE) shown in grey as a function of δ_{IOI}, as well as the binned data and the fitted softplus curve.

CHAPTER 6

CONCLUSION

In this thesis, I used neuroevolution to study the evolution of some of the most fundamental neural circuits such as 1) visual motion detection, 2) intraspecific collision avoidance using visual motion cues, 3) sound localization, and 4) event duration perception in rhythmic auditory stimuli. In particular, I used the Markov Brains platform that uses in silico Darwinian evolution, via a genetic algorithm (GA), to train neural networks that consist of binary neurons and are connected via logic gates. As explained in depth earlier, the circuit network, structure, and computation are all subject to evolution, which is an attempt to simulate how these neural circuits evolved in nature in the first place. This bottom-up approach is in contrast with more common methods used in computational neuroscience and artificial intelligence where researchers design rule-based systems, network structure, and its components. The evolutionary process and specific properties of the Markov Brains platform makes it a more plausible model of neural circuits in many respects.

The Markov Brains platform provides the possibility to explore the structure, complexity, and functionality of evolved neural circuits. For example, in chapters 2 and 4 I used a gate-knockout analysis to investigate the type of logic gates that are essential in evolved motion detection and sound localization circuits and I demonstrated the distribution of different types of logic gates that contribute to these neural circuits. In addition to analyzing the network structure and its components, it is also possible to test and analyze evolved agents in environments that are completely different from environments in which they evolved. This approach is particularly useful to isolate environmental factors that could play a role in the evolved behavior. In chapter 3 for example, I used a behavioral analysis in which the environmental factor under investigation was the apparent motion of the moving object (robot) in an agent's vision, namely regressive or progressive motion. Similarly in chapter 5, I evolved brains that can judge the duration of an auditory stimulus in a rhythmic sequence and then tested these evolved brains when exposed to out-of-rhythm oddball tones. Last but not least, the algorithms and computations of Markov Brains can be described in

terms of their state-space transitions. In chapter 5, for the first time I implemented a new technique that records a Markov Brain's neural activity as a sequence of transitions from one discrete state to another. In this type of analysis, a Markov Brain is represented as a finite state machine (FSM) which allows us to explore its state-space and analyze the brain's trajectories in the state-space when experiencing different stimuli in the environment, in order to discover algorithms and mechanisms behind its behavior.

In summary, I was able to utilize this powerful approach to address different questions and hypotheses regarding the fundamental neural circuits, the so called "widgets of intelligence". In what follows I briefly recapitulate some of these findings and also discuss the lessons I learned along the way in each project, and how they helped me make improvements in designing and conducting future research projects.

6.1 Visual Motion Detection

In chapter 2 I studied visual motion detection and found that evolution leads to a wide diversity of neuronal circuits even though each has the same function. I also observed that most circuits are more complex than one of the standard motion detection circuit models, the Reichardt detector, and showed that this increase in complexity is due to redundancy in the evolved circuits' structure. Measurements of mutational sensitivity showed that the evolved circuits were subject to additional selective pressures other than the basic functionality. But perhaps the most significant discovery in this project was that the wide diversity I observed in the evolution of Markov Brains performing motion detection was in accordance with patterns previously shown in the evolution of genetic circuits [194], functional systems based on biochemistry [200], as well as modeling and empirical studies of neuronal circuits with fixed wiring structure [152, 56]. This observation was the first stepping stone in establishing Markov Brains as a model system for the study of neural circuits evolved by Darwinian natural selection.

This study was also insightful for me in terms of experimental design decisions and research conduct. One of the examples of such design decisions concerned how to read the output neurons

for motion detection circuits. Initially, I tried a few common implementations: 1) assign three output neurons to each class, 2) assign two input neurons as outputs and read it as a two-bit binary value with possible outputs 00, 01, 10, 11. In the course of running the experiments I found that the aforementioned options have low evolvability especially because one of the classes (stationary object) is more common than the others (preferred direction and null direction). So I came up with a solution in which I assigned two different output patterns (01 and 10) to stationary objects. The reasoning behind this decision was that I considered the sum of the output values as the *firing rate* of the output neuron. Indeed, in the biological motion detection circuits of fruit flies, the motion state is encoded in terms of a neuron's firing rate. The other design decision was how to evaluate Markov Brains when seeing the visual input patterns. There were 16 possible input patterns and they fall into 3 different categories. Furthermore, their frequency distributions are not uniform; 10 of those patterns correspond to stationary objects, 3 of them are preferred motion and 3 are null direction. I tried two different approaches. First, I evaluated Markov brains with a fixed number of input patterns (for example 20) in which the probability distribution of different classes are uniform. Obviously, in this approach there are a lot of repetitions in evaluations of preferred direction (PD) and null direction (ND) classes. So I came up with a second solution in which I eliminated all the repetitions in evaluations, meaning I evaluate each agent with all possible 16 input patterns once, but I assigned different reward values to patterns that are more abundant. In other words, I constructed a non-uniform fitness function based on the non-uniform frequency of the three output classes. These two different approaches led to two different evolutionary outcomes but their differences were not significant for the results presented in [184]. These were all valuable lessons that shaped the experimental design in sound localization and time perception projects.

6.2 Intraspecific Collision-Avoidance Strategy based on Apparent Motion Cues

In chapter 3, I studied the intra-specific collision avoidance strategy based on apparent motion cues that was observed in *Drosophila melanogaster* [207, 33]. High-throughput data along with

mathematical analysis provided evidence for a strategy in which the apparent back-to-front motion (regressive motion) in a fly's retina is a cue to avoid collisions. I investigated possible selective pressures and environmental conditions for the evolution of this strategy. I showed that even though it is possible to evolve collision avoidance behavior in Markov Brains that uses regressive motion as the cue, it is highly unlikely that collision avoidance was the selective pressure behind the evolution of the observed behavior. The results of my evolutionary experiments clearly showed that the described behavior only evolves in a narrow range of experimental setups. Furthermore, I performed a mathematical analysis in which I calculated the probability of collisions in cases that generate an apparent regressive motion in a fly's retina. This analysis showed that in the experimental setup used in [207] only 20% of such events end up in collisions.

As discussed before, I managed to evolve Markov Brains that show a behavior similar to those observed in fruit flies. But it is worth mentioning that I tried a few different experimental setups and the explained behavior did not evolve in the beginning. First, I used a setup in which a group of flies were positioned in a two-dimensional arena and gained rewards for walking and incurred penalty for colliding with other flies (the fitness function described here is same as the one described in chapter 3). I also tested agents with and without the ability to make turns to avoid collisions. The observed behavior did not evolve in any of these setups. In particular, in the setup where agents had the ability to turn, agents evolved to circle around in a small space individually to avoid collisions while walking constantly (which was not surprising in hindsight). In a different setup, I put two flies in an arena without the ability to turn. The desired behavior did not evolve in this setup either. The optimal strategy that evolved here was that agents stopped once they sensed another fly regardless of the direction of the apparent motion. As a result, I recreated an experimental setup very similar to that used by [207] with a moving object that created a progressive or regressive motion in the agent's visual field. This setup also made me perform an analysis to calculate what percentage of the events that create a regressive motion in the retina result in collision. I believe the most valuable lesson learned for me in this project was to analyze and benefit from negative results, and also to start with simpler building blocks and make sure they work before proceeding

to a more complicated task. The latter lesson, i.e., building simpler components of a bigger system, was in fact the incentive behind the visual motion detection project.

6.3 Information Flow in Motion Detection and Sound Localization Circuits

In chapter 4, I studied whether transfer entropy (TE) measurement can accurately infer the flow of information in neural circuits, in particular, in motion detection and sound localization circuits. I addressed the question using different approaches. First, I calculated the accuracy of TE measurements in different types of logic gates and used their frequencies in neural circuits. Then, I used a different method in which I used TE as a proxy to infer information flow. In this approach, non-zero values of TE are equivalent to the existence of a causal relation between two neurons. I then generated the receiver operating characteristic (ROC) curves for each circuit. I also created the receptive fields and influence maps of each network using the connections and the logic used in the network in order to have a "ground-truth" model for information flow in the networks. These various approaches and analysis methods showed that the accuracy of TE measurements can be very sensitive to the type of circuit (the task it is performing), its connectivity structure, and its size. Furthermore, I showed that creating a ground-truth model can be a very hard task even if all the information about the network and its functionality is accessible. Finally, I showed that even in the absence of empirical limitations, inferring causation and information flow can be very challenging. For example, in our analysis we used neural recordings in the absence of any noise and we were able to record from every neuron in a neural circuit. Furthermore, we had access to the recordings of the brain for all possible sensory patterns. This just reminds us again that causality and identifying causal relations in a system is a very hard problem, as acknowledged before (see for example [145]). This problem becomes even harder when the subject matter is the nervous system, which arguably is the most complex system known to us. It also emphasized the fact that perhaps one of the missing components in the study of the brain is information theory, and probably one of the future breakthroughs in the field would follow the discovery of a new information-theoretic method that addresses causality.

6.4 Event Duration Perception in Rhythmic Auditory Stimuli

In chapter 5, I studied attentional entrainment as a model of event duration perception in rhythmic auditory stimuli. In particular, I tested two competing models of time perception in relation to attention, Scalar Expectancy Theory (SET) and Dynamic Attending Theory (DAT). In this project, I evolved Markov Brains that are able to perform duration judgment task in a rhythmic sequence of tones. These evolved brains can be considered as participants in a psychophysical experiment. We also performed psychometric tests that showed that these evolved brains have the same perceptual characteristics as human subjects [183]. For example, the discrimination threshold of evolved Markov Brains complies with Weber's law and furthermore, their point of subjective equality reveals similar trends to that of human subjects. I then tested the evolved brains against out-of-rhythm tones that they have not experienced during evolution. The psychometric results of these tests showed duration misperceptions that are similar to those experienced by human subjects.

In this project, I used a new method to analyze the computations and algorithms that Markov Brains used. I used the state-space transitions of Markov Brains that can reveal their computations, as well as how these brains pay attention to parts of the sensory input and do not pay attention to other parts. The results of this analysis showed that unlike what SET posits, the attention distributions in Markov Brains are not uniform in time at all. Furthermore, I observed that the attention distributions during the trials were also not in accordance with DAT, which predicts that attention peaks at the beginning of the rhythmic tones. Rather, evolved Markov Brains paid less attention to the beginning of the tones and their attention peaks coincided with the end point of the tone. These results suggest a new model of dynamic attending or attentional entrainment, where attention reaches its highest point when the stimulus is potentially the most *informative*, and attention drops when the stimulus is predictable. This new model can also be generalized to other modes of sensation such as visual attention. The generalized model would suggest that –similar to auditory attention in which attention peaks at specific time points– visual attention would be focused on those parts of the visual field that is predicted to be more informative. It would be interesting to design experiments that can test this new model of attention with human subjects, with

both auditory and visual sensory patterns. Aside from the proposed model of attention, the more important conclusion is the fact that we can test existing models of cognition using computational evolutionary methods and then we are able to suggest modification and even come up with new models. These new models make predictions that can be tested in biological brains.

To wrap up, I would like to suggest a possible future project that is in line with the research I did in this thesis, and in particular is inspired by the idea discussed before, namely that visual attention is focused on the most informative parts of an image. For this work, I propose to evolve Markov Brains that perform an image classification task via visual saccades that are driven by the information content (Shannon information [167]) of the image rather than the image saliency. The proposed project seems promising to me based on my experience in the field of computational cognitive science and using Markov Brains as the platform.

6.5 Information-Driven Image Classification via Saccadic Eye Movements

Here, I propose a project as a possible direction to pursue in the future, which involves "information-driven visual attention in image classification". This proposal is in part inspired by the work done by Olson et al. [142]. They conducted a series of experiments to evolve Markov brains that performed active image recognition of hand-written numerals in the MNIST dataset [103]. Unlike most widely used image recognition methods in which the classifier networks view the entire image and do not actively change the temporal or spatial structure of the data they receive, Olson et al. evolved classifiers that could view only a subset of the pixels in the image (a 3×3 sub-image) and could navigate which part of the image to view. As a result, the agents viewed a temporal sequence of sub-images rather than seeing the entire image at once. They were evolved to perform the image classification task by navigating and scanning the sub-images in a finite number of time steps. They ran 30 replicates of the evolutionary experiment, namely 30 populations, for nearly 250k generations and used only 1000 images from the MNIST training dataset (the original dataset consists of 60,000 images). They presented the results of their most successful run in which the agent with the highest performance only achieved 76% accuracy on testing dataset. The accu-

racy they achieved in their experiments is significantly lower than that of other machine learning methods such as K-nearest neighbors [88], support vector machines (SVM) [42], ANNs [126], and CNNs [36] which can be attributed to multiple factors I discuss here.

1. According to the data presented in their own paper [142], using a smaller portion of the dataset can result in lower accuracy. They presented results of training a decision tree model [147] on the smaller dataset, which only achieved 88.5% accuracy.

2. One of the configuration decisions they made in setting up the system was to put the agent in a random position in the image, which made the task much more difficult than it needed to be. The evidence for this speculation is the fact that in the early stages of the evolution the agent with the highest performance evolves to find the center point at the top edge of the image and uses it as a reference point in order to start its navigation through the image.

3. Another configuration factor that may have prevented the final performance from reaching higher levels is that the saccadic movements were limited to translating from a given position in the image to its neighboring points whereas biological saccadic eye movements enable transitioning from a point in the visual field to any other point.

Here, I propose modifications to the experimental setup for the evolution of image classification task via saccadic eye movements that addresses some of the issues discussed in [142]. I also suggest a different approach that attempts to improve performance results in the MNIST dataset. As mentioned before, the idea behind this approach relies mainly on navigating the visual attention or saccades based on the information content of the images. This captures how, for example, humans navigate their attention via saccades through salient regions of an image to recognize faces [75].

First, I investigate the entropy content of each pixel of the images in the MNIST dataset. The entropy for each pixel is calculated as:

$$H(X) = -\sum_{i} -p(x_i)\ \log(p(x_i)), \tag{6.1}$$

where x_i are the possible states a pixel can take on and the summation is over images of the dataset. Figure 6.1(A) shows the entropy content of pixels (in bits) for the images in the MNIST dataset. I used the same dataset used in [142], in which images were converted from greyscale to black and white; therefore, the possible states of a pixel are either 0 or 1.

Similarly, I can explore the entropy of variable C, which denotes the class to which an image belongs. In the MNIST dataset, variable C can take on values 0-9, i.e., c_i=0-9). Each class in the MNIST dataset has an equal number of images in the dataset, meaning there are 6000 images for each digit (there are 60,000 image in the MNIST dataset). As a consequence, the probability distribution of c_i is uniform, thus the entropy $H(C) = \log(10)$. Now we can link the entropy $H(C)$ to the entropy content of images and their pixels. For example, we can calculate how much entropy is reduced by looking at the value of a particular pixel. So we first calculate the conditional entropy $H(C|X)$ that shows the uncertainty in class variable C given the value of a particular pixel X:

$$H(C|X) = -\sum_{i,j} p(c_i, x_j)\, \log(p(c_i|x_j)), \tag{6.2}$$

Then, we can calculate how much the entropy of C can be reduced given the state of a particular pixel X:

$$I(C : X) = H(C) - H(C|X), \tag{6.3}$$

or the information shared between class variable C and a particular pixel X. Figure 6.1(B) shows the information $I(C : X)$ (in bits) shared between image classes and each pixel. For a uniform probability distribution the entropy is always maximal (the uncertainty is the highest) but as the entropy decreases, for example by subtracting a conditional entropy, the probability distribution becomes non-uniform (see [4]). In other words, $I(C : X)$ values in figure 6.1(B) show how the probability distribution of C can be distorted from a uniform distribution given the value of a single pixel. Figure 6.1(C) shows two different probability distributions of class variable C given the value of the pixel in the center (the pixel with the highest information in figure 6.1(B)) is either 0 or 1. Similarly, the entropy of C can be reduced further given the values of other pixels in the image, namely viewing one pixel of the image at a time. Furthermore, these results suggest that it

would be more efficient to first scan the pixels that reduce entropy the most, rather than scanning pixels in a random order. So we can design an experiment such that the agents saccade through the sequence of pixels with the highest information contents. Furthermore, in the proposed design the agents will see a group of adjacent pixels, for example a 2×2 sub-image at every time step instead of seeing one pixel at a time.

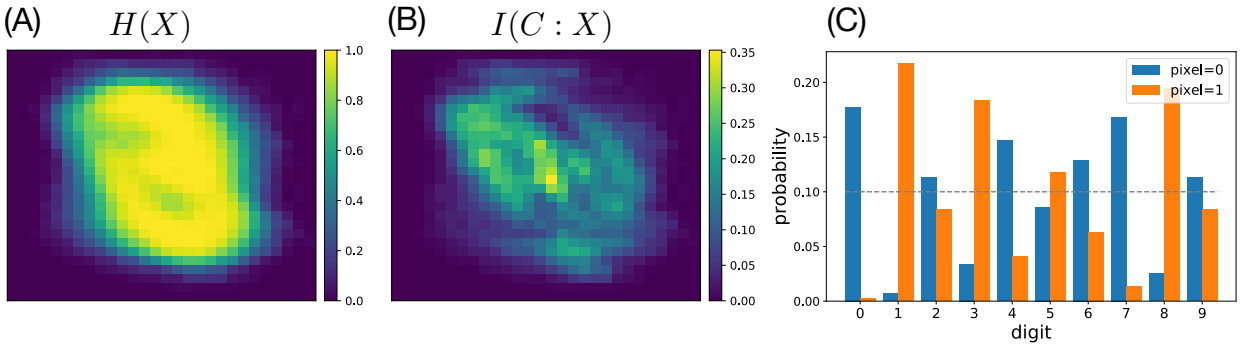

Figure 6.1: The images in the dataset are 28×28 pixels. (A) The entropy content (in bits) of MNIST dataset images per pixel, $H(X)$. (B) The information shared between each pixel and the class of the image, $I(C:X)$. (C) The probability distributions of class variable C given the pixel in the center is 0 or 1.

6.6 Experimental Setup

I evolve Markov Brains that see parts of an image through a 2×2 or 3×3 window at each time step and classify the image at the end of the sequence of saccades. The experimental setup proposed here is very similar to the experiments by [142] except that the saccade positions are predetermined by the results generated based on information content of the images in the dataset. More specifically, the agents will sense the sub-images at each position specified in the based on the information content and deliver their classification decision at the end.

6.6.1 Proof of concept

I also investigated whether the proposed approach can potentially improve the image classification performance for the MNIST dataset. To this end, I trained ANNs (multi-layered perceptrons, MLP)

that perform image classification on the MNIST dataset in which images that are fed to the network are partially masked. Masking of the images was performed based on two different criteria: 1) I masked images based on the entropy content of the pixels/sub-images across the dataset (see figure 6.1(A)), meaning the network only sees the sub-images with the high entropy contents, and 2) images are masked based on the entropy reduction in C given a sub-image (see figure 6.1(B)), where the network sees sub-images with high $I(C:X)$. Figure 6.2(A) and (B) shows the accuracy of the trained ANNs on masked images based on entropy of sub-images and the information shared between class C and sub-images, respectively. In these plots, the x-axis shows the thresholds by which the sub-images were masked, for example, the values of 0.2 on x-axis in figure 6.2(A) shows runs in which all the sub-images with $H(X)$ less than 0.2 were masked in the dataset. The y-axis on the left shows the accuracy of the network on the testing dataset and the y-axis on the right shows what percentage of the image was visible to the network. We observe that the network can still achieve an 80% accuracy on the testing dataset when only around 20% of the image is visible to the network when masking images based on entropy of sub-images, and they can achieve around 80% accuracy when only 15% of the image is visible based on $I(C:X)$. These results show that this approach has the potential to significantly improve experimental design and consequently the final results.

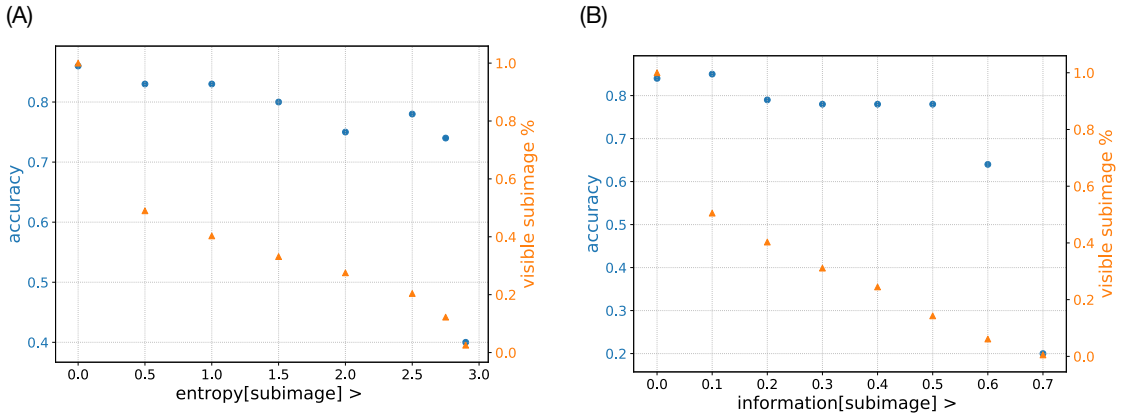

Figure 6.2: The performance of ANNs trained on masked images. Maskings were based on (A) the entropy content of sub-images in the dataset, and (B) the information shared between C and the sub-images.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Adami, C. Digital genetics: unravelling the genetic basis of evolution. *Nature Reviews Genetics 7* (2006), 109.

[2] Adami, C. What do robots dream of? *Science 314* (2006), 1093–1094.

[3] Adami, C. The use of information theory in evolutionary biology. *Annals of the New York Academy of Sciences 1256*, 1 (2012), 49–65.

[4] Adami, C. What is information? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374*, 2063 (2016), 20150230.

[5] Addyman, C., French, R. M., and Thomas, E. Computational models of interval timing. *Current Opinion in Behavioral Sciences 8* (2016), 140–146.

[6] Adelson, E. H., and Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A 2* (1985), 284–299.

[7] Ahn, Y. Y., Jeong, H., and Kim, B. J. Wiring cost in the organization of a biological neuronal network. *Physica A 367* (2006), 531–537.

[8] Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput Biol 10* (2014), e1003966.

[9] Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput Biol 10* (2014), e1003966.

[10] Albantakis, L., Marshall, W., Hoel, E., and Tononi, G. What caused what? a quantitative account of actual causation using dynamical causal networks. *Entropy 21*, 5 (2019), 459.

[11] Ay, N., and Polani, D. Information flows in causal networks. *Advances in complex systems 11*, 01 (2008), 17–41.

[12] Ayala, F. J., and Campbell, C. A. Frequency-dependent selection. *Ann Rev Ecol System 5* (1974), 115–138.

[13] Barlow, H., and Levick, W. R. The mechanism of directionally selective units in rabbit's retina. *J Physiol 178* (1965), 477–504.

[14] Barnett, L., Barrett, A. B., and Seth, A. K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys Rev Lett 103* (2009), 238701.

[15] Bates, D. M., and Watts, D. G. *Nonlinear Regression Analysis and its Applications*, vol. 2. Wiley New York, 1988.

[16] Beer, R. The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior 11* (2003), 209–243.

[17] Blount, Z. D., Borland, C. Z., and Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc Natl Acad Sci U S A 105* (2008), 7899–7906.

[18] Bolhuis, J. J., Brown, G. R., Richardson, R. C., and Laland, K. N. Darwin in mind: New opportunities for evolutionary psychology. *PLoS biology 9*, 7 (2011), e1001109.

[19] Bongard, J., Zykov, V., and Lipson, H. Resilient machines through continuous self-modeling. *Science 314* (2006), 1118–1121.

[20] Borst, A., and Egelhaaf, M. Principles of visual motion detection. *Trends Neurosci 12* (1989), 297–306.

[21] Borst, A., and Egelhaaf, M. Principles of visual motion detection. *Trends in neurosciences 12*, 8 (1989), 297–306.

[22] Borst, A., and Helmstaedter, M. Common circuit design in fly and mammalian motion vision. *Nat Neurosci 18* (2015), 1067.

[23] Bossomaier, T., Barnett, L., Harré, M., and Lizier, J. T. *An Introduction to Transfer Entropy*. Springer International, Cham, Switzerland, 2015.

[24] Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[25] Branson, K., Robie, A. A., Bender, J., Perona, P., and Dickinson, M. H. High-throughput ethomics in large groups of drosophila. *Nature Methods 6* (2009), 451–457.

[26] Buhusi, C. V., and Meck, W. H. What makes us tick? Functional and neural mechanisms of interval timing. *Nat Rev Neurosci 6* (2005), 755.

[27] Bunge, M. A. *Causality: The place of the causal principle in modern science*. Harvard University Press, Cambridge, MA, 1959.

[28] Buonomano, D. V., and Mauk, M. D. Neural network model of the cerebellum: temporal discrimination and the timing of motor responses. *Neural Computation 6* (1994), 38–55.

[29] Buzsáki, G. *Rhythms of the Brain*. Oxford University Press, New York, NY, 2006.

[30] C G, N., LaBar, T., Hintze, A., and Adami, C. Origin of life in a digital microcosm. *Phil Trans Roy Soc A 375* (2017), 20160350.

[31] Carnevale, N. T., and Hines, M. L. *The NEURON book*. Cambridge University Press, 2006.

[32] Casini, L., and Macar, F. Effects of attention manipulation on judgments of duration and of intensity in the visual modality. *Mem Cognit 25* (1997), 812–8.

[33] Chalupka, K., Dickinson, M., and Perona, P. Generalized regressive motion: A visual cue to collision. *arXiv preprint arXiv:1510.07573* (2015).

[34] Chapman, S., Knoester, D., Hintze, A., and Adami, C. Evolution of an artificial visual cortex for image recognition. In *Advances in Artificial Life, ECAL 12* (2013), pp. 1067–1074.

[35] Chesson, P. Mechanisms of maintenance of species diversity. *Ann Rev Ecol System 31* (2000), 343–366.

[36] Ciregan, D., Meier, U., and Schmidhuber, J. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition* (2012), IEEE, pp. 3642–3649.

[37] Coull, J. T., Vidal, F., Nazarian, B., and Macar, F. Functional anatomy of the attentional modulation of time estimation. *Science 303* (2004), 1506–1508.

[38] Cover, T. M., and Thomas, J. A. *Elements of Information Theory*. John Wiley, New York, NY, 1991.

[39] Cross, F. R., Buchler, N. E., and Skotheim, J. M. Evolution of networks and sequences in eukaryotic cell cycle control. *Phil Trans Roy Soc B 366* (2011), 3532–3544.

[40] Darwin, C. *The Descent of Man, and Selection in Relation to Sex*. John Murray, London, 1871.

[41] Darwin, C. *On the Origin of Species By Means of Natural Selection*. Murray, London, 1959.

[42] Decoste, D., and Schölkopf, B. Training invariant support vector machines. *Machine learning 46*, 1-3 (2002), 161–190.

[43] Duda, R. O., Hart, P. E., et al. *Pattern classification and scene analysis*, vol. 3. Wiley New York, 1973.

[44] Durstewitz, D. Self-organizing neural integrator predicts interval times through climbing activity. *Journal of Neuroscience 23* (2003), 5342–5353.

[45] Edlund, J. A., Chaumont, N., Hintze, A., Koch, C., Tononi, G., and Adami, C. Integrated information increases with fitness in the evolution of animats. *PLoS Comput Biol 7* (2011), e1002236.

[46] Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. A rotation and a translation suffice: Fooling cnns with simple transformations.

[47] Fechner, G. T. *Elemente der Psychophysik*, vol. 2. Breitkopf und Härtel, Leipzig, 1860.

[48] Floreano, D., Dürr, P., and Mattiussi, C. Neuroevolution: from architectures to learning. *Evolutionary intelligence 1*, 1 (2008), 47–62.

[49] Fogel, D. B., Fogel, L. J., and Porto, V. Evolving neural networks. *Biological cybernetics 63*, 6 (1990), 487–493.

[50] Fortuna, M. A., Zaman, L., Ofria, C., and Wagner, A. The genotype-phenotype map of an evolving digital organism. *PLoS Comput Biol 13* (2017), e1005414.

[51] Gauci, J., and Stanley, K. O. Autonomous evolution of topographic regularities in artificial neural networks. *Neural computation 22*, 7 (2010), 1860–1898.

[52] Getty, D. J. Discrimination of short temporal intervals: A comparison of two models. *Attention, Perception, & Psychophysics 18* (1975), 1–8.

[53] Gibbon, J. Scalar expectancy theory and Weber's law in animal timing. *Psychol. Rev. 84* (1977), 279–325.

[54] Gibbon, J., Church, R. M., and Meck, W. H. Scalar timing in memory. *Ann NY Acad Sci 423* (1984), 52.

[55] Gilbert, C. D., and Sigman, M. Brain states: Top-down influences in sensory processing. *Neuron 54* (2007), 677–96.

[56] Goaillard, J.-M., Taylor, A. L., Schulz, D. J., and Marder, E. Functional consequences of animal-to-animal variation in circuit parameters. *Nat Neurosci 12* (2009), 1424.

[57] González-González, A., Hug, S. M., Rodríguez-Verdugo, A., Patel, J. S., and Gaut, B. S. Adaptive mutations in RNA polymerase and the transcriptional terminator Rho have similar effects on escherichia coli gene expression. *Mol Biol Evol 34* (2017), 2839–2855.

[58] Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., and Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature 551* (2017), 45.

[59] Gould, S. J. *Wonderful Life: the Burgess Shale and the Nature of History*. WW Norton & Company, 1990.

[60] Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.

[61] Grondin, S. From physical time to the first and second moments of psychological time. *Psychological Bulletin 127* (2001), 22.

[62] Habenschuss, S., Jonke, Z., and Maass, W. Stochastic computations in cortical microcircuit models. *PLoS Comput Biol 9* (2013), e1003311.

[63] Halpern, J. Y. *Actual causality*. MiT Press, 2016.

[64] Hass, J., and Durstewitz, D. Neurocomputational models of time perception. In *Neurobiology of Interval Timing*, H. Merchant and V. de Lafuente, Eds. Springer, New York and Heidelberg, 2014, pp. 49–71.

[65] Hassenstein, B., and Reichardt, W. Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus. *Z Naturforsch B 11* (1956), 513–524.

[66] Hawkins, J., and Ahmad, S. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in neural circuits 10* (2016), 23.

[67] Hawkins, J., and Blakeslee, S. *On Intelligence*. Henry Holt and Co., New York, NY, 2004.

[68] Hilgetag, C. C., and Kaiser, M. Clustered organization of cortical connectivity. *Neuroinformatics 2* (2004), 353–60.

[69] Hintze, A., Edlund, J. A., Olson, R. S., Knoester, D. B., Schossau, J., Albantakis, L., Tehrani-Saleh, A., Kvam, P., Sheneman, L., Goldsby, H., et al. Markov brains: A technical introduction. *arXiv:1709.05601* (2017).

[70] Hintze, A., Kirkpatrick, D., and Adami, C. The structure of evolved representations across different substrates for artificial intelligence. *arXiv preprint arXiv:1804.01660* (2018).

[71] Hintze, A., and Miromeni, M. Evolution of autonomous hierarchy formation and maintenance. In *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems* (2014), pp. 366–367.

[72] Hopcroft, J. E., and Ullman, J. D. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Longman, Boston, MA, 1979.

[73] Hope, E. A., Amorosi, C. J., Miller, A. W., Dang, K., Heil, C. S., and Dunham, M. J. Experimental evolution reveals favored adaptive routes to cell aggregation in yeast. *Genetics 206* (2017), 1153–1167.

[74] Itti, L., and Koch, C. Computational modelling of visual attention. *Nat Rev Neurosci 2* (2001), 194.

[75] Itti, L., and Koch, C. Computational modelling of visual attention. *Nature reviews neuroscience 2*, 3 (2001), 194–203.

[76] James, R. G., Barnett, N., and Crutchfield, J. P. Information flows? A critique of transfer entropies. *Phys. Rev. Lett. 116* (2016), 238701.

[77] Janzing, D., Balduzzi, D., Grosse-Wentrup, M., Schölkopf, B., et al. Quantifying causal influences. *The Annals of Statistics 41*, 5 (2013), 2324–2358.

[78] Jaramillo, S., and Zador, A. M. The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nat Neurosci 14* (2011), 246–51.

[79] Jeffress, L. A. A place theory of sound localization. *Journal of comparative and physiological psychology 41*, 1 (1948), 35.

[80] Jo, J., and Bengio, Y. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561* (2017).

[81] Jones, M. R. Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychol Rev 83* (1976), 323–55.

[82] Jones, M. R., and Boltz, M. Dynamic attending and responses to time. *Psychol Rev 96* (1989), 459.

[83] Jones, M. R., Moynihan, H., MacKenzie, N., and Puente, J. Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychol Sci 13* (2002), 313–9.

[84] Joshi, N. J., Tononi, G., and C., K. The minimal complexity of adapting agents increases with fitness. *PLoS Comput Biol 9* (2013), e1003111.

[85] Juel, B. E., Comolatti, R., Tononi, G., and Albantakis, L. When is an action caused from within? quantifying the causal chain leading to actions in simulated agents. arXiv:1904.02995, 2019.

[86] Karmarkar, U. R., and Buonomano, D. V. Timing in the absence of clocks: encoding time in neural network states. *Neuron 53* (2007), 427–438.

[87] Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology 15* (2005), 1943–1947.

[88] Keysers, D., Deselaers, T., Gollan, C., and Ney, H. Deformation models for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 8 (2007), 1422–1435.

[89] Kimura, M., and Crow, J. F. The number of alleles that can be maintained in a finite population. *Genetics 49* (1964), 725–738.

[90] Kirkpatrick, D., and Hintze, A. Augmenting neuro-evolutionary adaptation with representations does not incur a speed accuracy trade-off. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (2019), pp. 177–178.

[91] Kirkpatrick, D., and Hintze, A. The role of ambient noise in the evolution of robust mental representations in cognitive systems. In *Artificial Life Conference Proceedings* (2019), MIT Press, pp. 432–439.

[92] Knill, D. C., and Richards, W. *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, Mass., 1996.

[93] Kriegeskorte, N., and Douglas, P. K. Cognitive computational neuroscience. *Nat Neurosci 21*, 9 (Sep 2018), 1148–1160.

[94] Kriegeskorte, N., and Douglas, P. K. Cognitive computational neuroscience. *Nat Neurosci 21* (2018), 1148.

[95] Kriegeskorte, N., and Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences 17* (2013), 401–412.

[96] Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience 2* (2008), 4.

[97] Kvam, P., Cesario, J., Schossau, J., Eisthen, H., and Hintze, A. Computational evolution of decision-making strategies. In *Proceedings 37th Annual Meeting of the Cognitive Science Society* (Austin, TX, 2015), Noelle, D. C. et al., Ed., Cognitive Science Society.

[98] Kvam, P., Cesario, J., Schossau, J., Eisthen, H., and Hintze, A. Computational evolution of decision-making strategies. In *Proc. 37th Annual Conf. of the Cognitive Science Society*, D. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. Jennings, and P. P. Maglio, Eds. Cognitive Science Society, Austin, TX, 2015, pp. 1225–1230.

[99] LaBar, T., and Adami, C. Evolution of drift robustness in small populations. *Nature Comm 8* (2017), 1012.

[100] LaBar, T., Hintze, A., and Adami, C. Evolvability tradeoffs in emergent digital replicators. *Artificial Life 22* (2016), 483–498.

[101] Large, E. W., and Jones, M. R. The dynamics of attending: How people track time-varying events. *Psychol. Rev. 106* (1999), 119 – 159.

[102] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature 521*, 7553 (2015), 436–444.

[103] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[104] Lenski, R. E., Ofria, C., Pennock, R. T., and Adami, C. The evolutionary origin of complex features. *Nature 423*, 6936 (2003), 139–144.

[105] Lewontin, R. C., and Hubby, J. L. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura . *Genetics 54* (1966), 595–609.

[106] Lind, P. A., Farr, A. D., and Rainey, P. B. Experimental evolution reveals hidden diversity in evolutionary pathways. *Elife 4* (2015), e07074.

[107] Lizier, J. T., and Prokopenko, M. Differentiating information transfer and causal effect. *The European Physical Journal B 73*, 4 (2010), 605–615.

[108] Macar, F., Grondin, S., and Casini, L. Controlled attention sharing influences time estimation. *Mem Cognit 22* (1994), 673–86.

[109] Macmillan, N. A., and Creelman, C. D. *Detection theory: A user's guide*. Psychology press, 2004.

[110] Maloney, E. S. *Chapman Piloting, Seamanship and Small Boat Handling*. Hearst Marine Books, 1989.

[111] Marder, E. Variability, compensation, and modulation in neurons and circuits. *Proc Natl Acad Sci U S A 108*, Suppl 3 (2011), 15542–15548.

[112] Markram, H. The blue brain project. *Nature Reviews Neuroscience 7*, 2 (2006), 153–160.

[113] Marr, D., and Ullman, S. Directional selectivity and its use in early visual processing. *Proc R Soc Lond B 211* (1981), 151–180.

[114] Marstaller, L., Hintze, A., and Adami, C. The evolution of representation in simple cognitive networks. *Neural computation 25*, 8 (2013), 2079–2107.

[115] Marstaller, L., Hintze, A., and Adami, C. The evolution of representation in simple cognitive networks. *Neural Comput 25* (2013), 2079–2107.

[116] Marstaller, L., Hintze, A., and Adami, C. The evolution of representation in simple cognitive networks. *Neural Comput 25* (2013), 2079–2107.

[117] Matell, M. S., and Meck, W. H. Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cognitive Brain Research 21* (2004), 139–170.

[118] Matthews, W. J., and Meck, W. H. Temporal cognition: Connecting subjective time to perception, attention, and memory. *Psych Bull 142* (2016), 865.

[119] Mazzucato, L., La Camera, G., and Fontanini, A. Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. *Nat Neurosci 22* (2019), 787–796.

[120] McAuley, J. D. *Perception of time as phase: Toward an adaptive-oscillator model of rhythmic pattern processing*. PhD thesis, Indiana University, Indianapolis, IN, 1995.

[121] McAuley, J. D., and Fromboluti, E. K. Attentional entrainment and perceived event duration. *Philos Trans R Soc Lond B Biol Sci 369* (2014), 20130401.

[122] McAuley, J. D., and Jones, M. R. Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. *J Exp Psychol Hum Percept Perform 29* (2003), 1102–25.

[123] McAuley, J. D., Jones, M. R., Holub, S., Johnston, H. M., and Miller, N. S. The time of our lives: Life span development of timing and event tracking. *J Exp Psychol Gen 135* (2006), 348–67.

[124] McAuley, J. D., and Kidd, G. R. Effect of deviations from temporal expectations on tempo discrimination of isochronous tone sequences. *J Exp Psychol Hum Percept Perform 24* (1998), 1786–800.

[125] McFarland, D. J. Decision making in animals. *Nature 269* (1977), 15–21.

[126] Meier, U., Ciresan, D. C., Gambardella, L. M., and Schmidhuber, J. Better digit recognition with a committee of simple neural nets. In *2011 International Conference on Document Analysis and Recognition* (2011), IEEE, pp. 1250–1254.

[127] Michalewicz, Z. *Genetic Algorithms + Data Strucures = Evolution Programs*. Springer Verlag, New York, 1996.

[128] Middlebrooks, J. C., and Green, D. M. Sound localization by human listeners. *Annual review of psychology 42*, 1 (1991), 135–159.

[129] Miller, J. E., Carlson, L. A., and McAuley, J. D. When what you hear influences when you see: Listening to an auditory rhythm influences the temporal allocation of visual attention. *Psychol Sci 24* (2013), 11–8.

[130] Moore, B. C. *An introduction to the psychology of hearing*. Brill, 2012.

[131] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2574–2582.

[132] Moray, N. Where is capacity limited? a survey and a model. *Acta Psychologica 27* (1967), 84–92.

[133] Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 427–436.

[134] Nobre, A. C., Correa, A., and Coull, J. T. The hazards of time. *Curr Opin Neurobiol 17* (2007), 465–70.

[135] Nosil, P. *Ecological Speciation*. Oxford University Press, Oxford (UK), 2012.

[136] Oizumi, M., Albantakis, L., and Tononi, G. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Comput Biol 10* (2014), e1003588.

[137] Olson, R. S., Haley, P. B., Dyer, F. C., and Adami, C. Exploring the evolution of a trade-off between vigilance and foraging in group-living organisms. *Royal Society open science 2*, 9 (2015), 150135.

[138] Olson, R. S., Haley, P. B., Dyer, F. C., and Adami, C. Exploring the evolution of a trade-off between vigilance and foraging in group-living organisms. *R Soc Open Sci 2* (2015), 150135.

[139] Olson, R. S., Hintze, A., Dyer, F. C., Knoester, D. B., and Adami, C. Predator confusion is sufficient to evolve swarming behaviour. *J R Soc Interface 10* (2013), 20130305.

[140] Olson, R. S., Hintze, A., Dyer, F. C., Knoester, D. B., and Adami, C. Predator confusion is sufficient to evolve swarming behaviour. *Journal of The Royal Society Interface 10* (2013), 20130305.

[141] Olson, R. S., Knoester, D. B., and Adami, C. Critical interplay between density-dependent predation and evolution of the selfish herd. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation* (2013), pp. 247–254.

[142] Olson, R. S., Moore, J. H., and Adami, C. Evolution of active categorical image classification via saccadic eye movement. In *International Conference on Parallel Problem Solving from Nature* (2016), Springer, pp. 581–590.

[143] Palmer, S. E. *Vision science: Photons to Phenomenology*. MIT Press, Cambridge, MA, 1999.

[144] Paul, L. A., Hall, N., and Hall, E. J. *Causation: A user's guide*. Oxford University Press, 2013.

[145] Pearl, J. *Causality: models, reasoning and inference*, vol. 29. Springer, 2000.

[146] Pearl, J. *Causality*. Cambridge University Press, 2009.

[147] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research 12* (2011), 2825–2830.

[148] Phillips, W. A., and Singer, W. In search of common foundations for cortical computation. *Behavioral and Brain Sciences 20*, 4 (1997), 657–683.

[149] Pickles, J. *An introduction to the physiology of hearing*. Brill, 2013.

[150] Poirazi, P., Brannon, T., and Mel, B. W. Pyramidal neuron as two-layer neural network. *Neuron 37*, 6 (2003), 989–999.

[151] Polsky, A., Mel, B. W., and Schiller, J. Computational subunits in thin dendrites of pyramidal cells. *Nature neuroscience 7*, 6 (2004), 621–627.

[152] Prinz, A. A., Bucher, D., and Marder, E. Similar network activity from disparate circuit parameters. *Nat Neurosci 7* (2004), 1345.

[153] Proust, M. *A la Recherche du Temp Perdu*. Gallimard, Nouvelle Revue Française, Paris, 1919-1927.

[154] Qian, J., Hintze, A., and Adami, C. Colored motifs reveal computational building blocks in the C. elegans brain. *PLoS ONE 6* (2011), e17013.

[155] Raftery, A. E. Bayesian model selection in social research. *Sociological Methodology 25* (1995), 111–164.

[156] Rainey, P. B., Buckling, A., Kassen, R., and Travisano, M. The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends Ecol Evol 15* (2000), 243–247.

[157] Raman, K., and Wagner, A. The evolvability of programmable hardware. *J R Soc Interface 8* (2010), 269–281.

[158] Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision* (2016), Springer, pp. 525–542.

[159] Rich, E. L., and Wallis, J. D. Decoding subjective decisions from orbitofrontal cortex. *Nat Neurosci 19* (2016), 973–980.

[160] Richelle, M., Lejeune, H., Defays, D., Greenwood, P., Macar, F., and Mantanus, H. *Time in Animal Behaviour*. Pergamon Press, New York, NY, 2013.

[161] Rivoire, O., and Leibler, S. The value of information for populations in varying environments. *Journal of Statistical Physics 142*, 6 (2011), 1124–1166.

[162] Rosenzweig, M. L. *Species Diversity in Space and Time*. Cambridge University Press, Cambridge (UK), 1995.

[163] Schiff, W., Caviness, J. A., and Gibson, J. J. Persistent fear responses in rhesus monkeys to the optical stimulus of “looming". *Science 136* (1962), 982–3.

[164] Schossau, J., Adami, C., and Hintze, A. Information-theoretic neuro-correlates boost evolution of cognitive systems. *Entropy 18*, 1 (2015), 6.

[165] Schreiber, T. Measuring information transfer. *Physical Review Letters 85*, 2 (2000), 461.

[166] Seidemann, E., Meilijson, I., Abeles, M., Bergman, H., and Vaadia, E. Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. *J Neurosci 16* (1996), 752–68.

[167] Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal 27*, 3 (1948), 379–423.

[168] Shannon, C. E. Communication theory of secrecy systems. *Bell system technical journal 28*, 4 (1949), 656–715.

[169] Sheneman, L., and Hintze, A. Evolving autonomous learning in cognitive networks. *Scientific reports 7*, 1 (2017), 1–11.

[170] Sheneman, L., Schossau, J., and Hintze, A. The evolution of neuroplasticity and the effect on integrated information. *Entropy 21*, 5 (2019), 524.

[171] Shomrat, T., Graindorge, N., Bellanger, C., Fiorito, G., Loewenstein, Y., and Hochner, B. Alternative sites of synaptic plasticity in two homologous “fan-out fan-in" learning and memory networks. *Curr Biol 21* (2011), 1773–1782.

[172] Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., Boly, M., Postle, B. R., and Tononi, G. The neural correlates of dreaming. *Nature Neuroscience 20* (2017), 872.

[173] Sloss, A. N., and Gustafson, S. 2019 evolutionary algorithms review. *Genetic Programming Theory and Practice XVII* (2020), 307.

[174] Sorrells, T. R., Booth, L. N., Tuch, B. B., and Johnson, A. D. Intersecting transcription networks constrain gene regulatory evolution. *Nature 523* (2015), 361.

[175] Sporns, O. *Networks of the Brain*. MIT Press, Cambridge, MA, 2011.

[176] Stanley, K. O., and Miikkulainen, R. Evolving neural networks through augmenting topologies. *Evolutionary computation 10*, 2 (2002), 99–127.

[177] Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., and Bialek, W. Entropy and information in neural spike trains. *Phys Rev Lett 80* (1998), 197.

[178] Su, J., Vargas, D. V., and Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation 23*, 5 (2019), 828–841.

[179] Sun, J., and Bollt, E. M. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena 267* (2014), 49–57.

[180] Sved, J. A., Reed, T. E., and Bodmer, W. F. The number of balanced polymorphisms that can be maintained in a natural population. *Genetics 55* (1967), 469–481.

[181] Taylor, M. B., Phan, J., Lee, J. T., McCadden, M., and Ehrenreich, I. M. Diverse genetic architectures lead to the same cryptic phenotype in a yeast cross. *Nature Comm 7* (2016), 11669.

[182] Tehrani-Saleh, A., and Adami, C. Can transfer entropy infer information flow in neuronal circuits for cognitive processing? *Entropy 22*, 4 (2020), 385.

[183] Tehrani-Saleh, A., and Adami, C. Psychophysical tests reveal that evolved artificial brains perceive time like humans. In *ALIFE 2021: The 2021 Conference on Artificial Life* (2021), MIT Press.

[184] Tehrani-Saleh, A., LaBar, T., and Adami, C. Evolution leads to a diversity of motion-detection neuronal circuits. In *Proceedings of Artificial Life 16* (Cambridge, MA, 2018), T. Ikegami, N. Virgo, O. Witkowski, M. Oka, R. Suzuki, and H. Iizuka, Eds., MIT Press, pp. 625–632.

[185] Tehrani-Saleh, A., McAuley, J. D., and Adami, C. Mechanism of perceived duration in artificial brains suggests new model of attentional entrainment. *bioRxiv* (2019), 870535.

[186] Tehrani-Saleh, A., Olson, R., and Adami, C. Flies as ship captains? Digital evolution unravels selective pressures to avoid collision in Drosophila. In *Proc. Artificial Life 15* (Cambridge, MA, 2016), C. G. et al., Ed., MIT Press.

[187] Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., and Gaut, B. S. The molecular diversity of adaptive convergence. *Science 335* (2012), 457–461.

[188] Thomas, E., and Weaver, W. Cognitive processing and time perception. *Atten. Oercept. Psychophys. 17* (1975), 363–367.

[189] Tononi, G., Boly, M., Massimini, M., and Koch, C. Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience 17*, 7 (2016), 450.

[190] Tononi, G., and Koch, C. Consciousness: here, there and everywhere? *Phil. Trans. R. Soc. B 370* (2015), 20140167.

[191] Treisman, M. Temporal discrimination and the indifference interval. implications for a model of the “internal clock". *Psychol Monogr 77* (1963), 1–31.

[192] Treue, S., and Martínez Trujillo, J. C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature 399* (1999), 575–9.

[193] Tse, P. U., Intriligator, J., Rivest, J., and Cavanagh, P. Attention and the subjective expansion of time. *Percept Psychophys 66* (2004), 1171–89.

[194] Tsong, A. E., Miller, M. G., Raisner, R. M., and Johnson, A. D. Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell 115* (2003), 389–399.

[195] Tsong, A. E., Tuch, B. B., Li, H., and Johnson, A. D. Evolution of alternative transcriptional circuits with identical logic. *Nature 443* (2006), 415.

[196] Turing, A. M. Computing machinery and intelligence. *Mind 59*, 236 (1950), 433–460.

[197] VanRullen, R., and Koch, C. Is perception discrete or continuous? *Trends in Cognitive Sciences 7* (2003), 207–213.

[198] VanRullen, R., Reddy, L., and Koch, C. Attention-driven discrete sampling of motion perception. *Proc. Natl. Acad. Sci. U.S.A. 102* (2005), 5291–5296.

[199] Vicente, R., Wibral, M., Lindner, M., and Pipa, G. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational Neuroscience 30*, 1 (2011), 45–67.

[200] Wagner, A. *The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems*. Oxford University Press, Oxford (UK), 2011.

[201] Wibral, M., Lizier, J. T., and Priesemann, V. Bits from brains for biologically inspired computing. *Frontiers in Robotics and AI 2* (2015), 5.

[202] Wibral, M., Vicente, R., and Lindner, M. Transfer entropy in neuroscience. In *Directed Information Measures in Neuroscience*. Springer, 2014, pp. 3–36.

[203] Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature 412* (2001), 331.

[204] Williams, P. L., and Beer, R. D. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* (2010).

[205] Yao, X., and Liu, Y. A new evolutionary system for evolving artificial neural networks. *IEEE transactions on neural networks 8*, 3 (1997), 694–713.

[206] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792* (2014).

[207] Zabala, F., Polidoro, P., Robie, A., Branson, K., Perona, P., and Dickinson, M. H. A simple strategy for detecting moving objects during locomotion revealed by animal-robot interactions. *Current Biology 22* (2012), 1344–1350.

[208] Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications 10* (2019), 3770.

[209] Zhang, C., Liao, Q., Rakhlin, A., Sridharan, K., Miranda, B., Golowich, N., and Poggio, T. Theory of Deep Learning IIb: Generalization properties of SGD. Tech. Rep. arXiv:1801.02254, Center for Brains, Minds, and Machines, 2018.