Φ

Recent Topics in Machine Consciousness and the Evolution of Animats With Simulated Consciousness and Collective Behavior

University Magdeburg
Institute for Intelligent Systems

Master Thesis in Data- and Knowledge-Engineering

Dominik Fischer (B. Sc.)
Student-ID: **210310**

Submitted on the 12th April 2017

1^{st} Examiner: Prof. Dr. Sanaz Mostaghim
2^{nd} Examiner: Larissa Albantakis (Ph.D.)

Contents

1 Introduction **1**

1.1 Reasons to Delve Into MC 2

1.2 The Three Major Objectives 3

1.3 Scientific Demarcation 4

1.4 Structural Overview 5

1.5 How to Understand Consciousness 5

2 Overview on Modern Studies of Consciousness and MC **7**

2.1 Assessment of Human-Like Intelligence in Machines 7

2.2 On the Hard and Soft Problem of Consciousness 8

2.3 Consciousness is Not Provable and Unlikely to Reproduce 9

2.4 Modern Approaches on Consciousness Studies 10

2.5 Global Workspace Theory of Consciousness and LIDA 11

2.6 Further Readings on Machine Consciousness 14

3 Background on Consciousness Studies **16**

3.1 The Maze of Consciousness is Solvable 16

3.2 The Integrated Information Theory Makes Consciousness Testable 19

3.2.1 Information and Integration Explained 20

3.2.2 Postulates and Axioms of Consciousness 21

3.2.3 Measuring Integrated Information 23

3.2.4 Supports and Critics of IIT 24

3.2.5 Animats with Integrated Information 26

3.2.6 Alternatives to Calculate Integrated Information 26

3.3 Collective Behavior Through Evolutionary Algorithms 27

3.4 Markov Brains as Finite State-Machines 28

3.5 Can a Machine Develop Real Consciousness? 30

4 Model to Evolve Conscious Animats with Collective Behavior **31**

4.1 A 2D-Matrix Environment 31

4.2 Six Different Animat Designs 32

4.3 Fitness Function To Evolve Collective Behavior 34

4.4 The Communication and Cooperation of Animats 34

4.5 The Workflow of the Evaluation Procedure 35

4.6 Required Data Analysis 35

5 Implementation of the Evolution Model 39

5.1 Evolving Animats with Integrated Information 39

5.2 Calculating Integrated Information for Elite Organisms 44

6 Results and Evaluation of the Experiments 46

6.1 Experimental Setup 46

6.2 The Evolution of Fitness Compared to Integrated Information . 47

6.3 Further Insights on Behavior and Integration 51

6.4 On the Lack of Correlation Between Fitness and Measures of Integrated Information 55

6.5 Dissect of the Best Animat Brain 56

7 Future Work and Discussion 58

8 Conclusion 59

A Appendix I

A.1 Scripts Used for the Experiments I

A.2 Quality of Evaluations II

A.3 All Line of Decents III

A.4 Further Heat Maps of all Best Performers VII

B Literature XI

C List of Figures XVIII

D List of Tables XIX

E Glossary XX

Abstract

Studying consciousness has increased in the past two decades. Artificial intelligence is improving year by year and solves more and more complex tasks. Achieving machines with human-like intelligence, constructs a further question. People will ask if such machines do not only have human-like intelligence but also human-like consciousness or are such machines still unconscious? On the other side, there might be already machines, which are capable of doing this. How can we prove that an artificial machine has conscious experience? This work provides an overview of the most discussed approaches about consciousness and machine consciousness. In order to investigate how consciousness develops during the evolution of collective systems with intelligent behavior, this work provides a model and implementation to simulate cooperating artificial intelligent agents (animats). The level of integrated information, which is a measure of consciousness according to the integrated information theory, is observed over several stages in the evolution using the integrated information theory. It is also discussed what the main differences, between the behavior of high-conscious and low-conscious agents in a group of animats, are.

Acknowledgements

For the great support with major questions and problems, I want to shout out special thanks to people, which supported me without prior knowledge about my thesis. Larissa Albantakis, from the School of Medicine and Public Health at the University of Wisconsin-Madison, helped me understanding the principles of the integrated information theory and helped me during the validation of the calculations of integrated information. Furthermore, I want to thank Arend Hintze and Clifford Bohm, from the Evolving Artificial Intelligence Laboratory at the Michigan State University, giving valuable recommendations during the development of the extension of the evolution framework used in this work.

1 Introduction

Already about seventy years ago Alan Turing was pondering the question “Can machines think?” [Tur50]. And already in the second sentence of his work he pointed out that it is important to write down the meaning of the word *machine* and the word *think*. Now, seventy years later, variations of this question are still discussed, more than ever [TK15; ZD14; TK14; Bla12; SP11; Leg08; Hol07; FP06; ON01; Fra00; Bos98; Sea97; Sea90; Baa97]. While many researchers in machine learning simply try to pass the Turing test with masterpieces of intelligent robots, there is a marginal but growing group intends to go a step further. They try to find out, if it would be possible for machines to develop some sort of inner life or, with the terminology of this work, if a machine could be conscious. More or less it is the same question Turing had, but the majority explained the word *thinking* only by something like *“...the ability to solve problems only humans could solve...”*. In this work thinking should be understood in a sense of *the ability to experience* the world, there will be a detailed explanation later.

Mankind developed machines, which makes it obvious that a definition of the subject *machine* is not the problem. A clear challenge is to find a definition of consciousness [Gel14] and it is even harder to develop artificial conscious machines. Three major questions have to be answered to solve the puzzle:

1. What is the unified definition of consciousness?
2. How can we test consciousness?
3. How to implement consciousness in a machine?
4. What are the benefits and harms of living with conscious machines?

In the field of *Machine Consciousness (MC)*, there are many different interpretations of how to model MC. In most models, consciousness is seen as a software module, which gives rise to special conscious behavior [Ann16]. These models are good to understand basics and problems of consciousness but are not working on phenomenal consciousness, the ability to experience the world [Cha95].

In the past decade, a novel theory about consciousness called *Integrated Information Theory (IIT)* was proposed, revised and extended, which is a research program led by Prof. Giulio Tononi [Ton04; Ton12; OAT14]. The theory defines consciousness as the result of a highly integrated system of elements or mechanisms (e.g. neurons in the brain). The main advantage of this theory, from the computer science perspective is, that it provides measures for quality and quantity of consciousness. This means that IIT is currently the only known theory, which makes consciousness testable.

There are already works successfully demonstrating that simulated adaptive systems increase their level of integrated information over the course of evolution [Alb+14; Edl+11]. In such works, the simulated machines (called *animats*, ref. section 3.2.5) are only interacting alone in the environment. There are no investigations on groups of animats and whether they would evolve integrated information. Therefore a goal of this work is to find a model and implementation to simulate artificial collective animats. It is expected that depending on the architecture of the animats and depending on the swarm size of interacting animats, not only the performance and behavior of an animat will change, but also the quantity of integrated information. It should be investigated what could influence the development of integrated information in an individual organism and how the knowledge about neighbors and the amount of neighbors can influence consciousness and develop swarm behavior.

In the following section, there is a deeper explanation about the motivation and how exactly the objectives of this work are defined. This is followed by the scientific demarcation and the structural overview of this work. In the end of this chapter, we introduce an explanation on how to understand the term *consciousness* in this work.

1.1 Reasons to Delve Into MC

Machines solve more and more complex tasks, tasks we thought only humans are able to solve. They pass the Turing test more often and even more trust is given to artificial intelligence than ever before (e. g. the self-driving car project Waymo by Google[1] or IBM's super-computer Watson[2]). By common sense, soon it will not be possible to distinguish if a system we are communicating with has a conscious mind or not. This can lead to serious ethical conflicts. That is why it is important to investigate measures of consciousness for machines. They would help not only to determine if a machine is conscious or not but might also help to develop more advanced machines.

From the neurosciences' point of view simulating conscious populations has also another motivation. Consciousness in science is a vaguely defined subject. It is still a lot of research necessary to understand consciousness better. Therefore doing simulations, where it is possible to experiment with

[1] `https://www.google.com/selfdrivingcar/`

[2] `https://www.ibm.com/watson/`

the internal structures of organisms as well as the structure of the environment, could accelerate the research in the field of consciousness [Reg13]. Simulating is also a cheap and fast way to test theories or models, which should also hold for simulations of conscious organisms.

Furthermore, machine consciousness is not only a topic in the world of artificial intelligence or neurosciences. Even the philosophic world discusses if it is really possible for machines to achieve consciousness or if they already have one [Sea93; Tur50; Bla03]. Most of the philosophical theories about machine consciousness cannot be falsified until we are finally able to show what gives rise to consciousness in an organism and if we are able to put it in a machine. This means that it can also be supporting for the philosopher if there are implementations of simulated consciousness.

1.2 The Three Major Objectives

The goal of this work is to develop a vertical proceeding, which starts from philosophical and theoretical approaches and ends in modeling and implementing a lightweight simulation of collective animats. The tasks can be divided into three different stages:

Stage One: Overview on Machine Consciousness

Research in machine consciousness is rather new. There is a variety of different theories, interpretations and approaches out of many different scopes, like computer science, philosophy and cognitive sciences. That is why it is important to give an overview of the most significant works in this field.

As it is argued later, this work applies IIT. This theory requires discrete dynamical systems of interacting elements, which are then able to give rise to consciousness [Ton04]. Due to the motivation of this work, these systems should interact simultaneously in the same environment and evolve collective behavior.

Stage Two: Modeling of State-Based Animats with Collective Behavior

A model needs to be developed where it is possible to evolve a set of state-based animats with the constraint that they interact together in the same environment. It is also necessary to measure integrated information in animats with collective behavior.

The developed model should also be implemented. Furthermore different architectures and experiment settings should be tested to observe influential factors regarding the performance of the animats and their integrated information.

Stage Three: Implementation of Simulated Collective Machine Consciousness

Given the simulation model, the performance and integrated information of animats should vary in different experiment settings. Statements about the quality of consciousness of intelligent animats with collective behavior could be made with such observations.

In the real world, collective behavior is observable in swarms of animals: birds flying around, ants on the search for food or wolfs hunting a deer. In a group, such organisms are able to achieve better results, instead of working alone [Rei+15]. In other species where cooperation is necessary to increase the chance to stay alive during an attack, e.g. swarms of anchovies while a shark is attacking. From the outside perspective, this collective behavior can be seen to be intelligent and optimization algorithms are formed using this inspiration. In this work, animats should be modeled in an environment where they have to work together to stay alive. The question is, how they perform with different kinds of internal structures. How is the evolution of integrated information related to the organism's performance? This could be helpful for the development of new animats, acting in unfamiliar environments and also could support the IIT.

1.3 Scientific Demarcation

This work will not develop a new model of consciousness or of consciousness in machines. The goal is to use a widespread abstract theory, which is implementable in computer simulations. It should not just be assumed that simulated animats are conscious but also tested.

According to IIT, the simulated agents will not have real consciousness, while they have a high level of integrated information. Even if the agents would be physical robots, controlled by a simulated brain, they would not be conscious. Only a microphysical instantiation of an animat with integrated information, not a simulation, would be conscious according to IIT. Furthermore, if there is a supercomputer who simulates a perfect black hole, still this black hole exists only in the simulation. According to the IIT, conscious machines are not impossible, but they would require a new physical architecture of the computer as we know today (a von Neumann architectured robot would not be conscious) [TK15].

The work implements not a nature inspired collective behavior. It is seen as a better approach to implement a pretty small-scale model, which has the ability to grow in future research rather than pondering with the challenging task to find a solution for complex animal behavior.

1.4 Structural Overview

Since this thesis consists of three different objectives a short overview of the structure is given. First, there will be an interdisciplinary overview to machine consciousness as well as major western approaches on consciousness and how it can be seen in computer science. It will also be shown how to categorize works in this field.

Second, a broad background is given to the IIT as well as the supporting philosophy behind. The theory builds a bridge between clear axioms on conscious beings and computable postulates on physical organisms [OAT14]. There is also a short introduction to the evolution of intelligent systems in computer science.

Third, a simulation model, for social animats performing a collective behavior, is developed. In this part, not only the architecture of the animats is shown but additionally the environmental design, the workflow as well as the performance evaluation. Afterwards, the model is implemented using a framework called *MABE*[3] and evaluated with the *pyphi*[4] framework.

In the evaluation section, the experiment results are described and statements about the influential factors of the rise of consciousness in animats are made. Additionally, the collective behavior and correlation between the performance of the animats are investigated. This work ends with a discussion about the model, implementation and evaluation as well as with a prediction on future works.

1.5 How to Understand Consciousness

“What is consciousness?” – Since centuries or even thousands of years scientists, theologists and of course spiritual seekers are working on this question. There can be found many different definitions and interpretations, also strongly

[3] `https://github.com/ahnt/MABE`

[4] `https://github.com/wmayner/pyphi`

dependent on the field of studies. That is why we have to strongly isolate the concept of consciousness for our work. Therefore statements of modern western philosophers are used, which also supports the IIT.

John Searle sees consciousness as a biological phenomenon. He states that if it is possible to understand completely how our brain works, we are also able to find a definition for consciousness and how it would be possible to rebuild it [Sea93]. Consciousness has nothing to do with self-consciousness, knowledge or something like attention, but is more the state of sentience and awareness which someone has while dreaming or is awake (not in a dreamless sleep). To keep this short, Searle's interpretations of consciousness could be seen as the ability to make experiences.

> *"First, I want to argue that we simply know as a matter of fact that brain processes cause conscious states. We don't know the details about how it works and it may well be a long time before we understand the details involved. Furthermore, it seems to me an understanding of how exactly brain processes cause conscious states may require a revolution in neurobiology." [Sea93]*

Next, it is important to define how machine consciousness is seen in this work. There are two possible categories in a coarse-grained level [Hol03]:

1. **Weak Artificial Consciousness:** Machines or systems that only simulate processes, that are correlated to consciousness or what we think consciousness is.

2. **Strong Artificial Consciousness:** These are machines or systems, which are conscious of their being. As you know that you can experience this moment, the machine would have some sort of the same experience.

According to IIT animats with integrated information would be conscious if they would exist as a physical organism. Such organisms would have strong artificial consciousness. Since animats in this work are only simulated organisms, they have only weak artificial consciousness. Until we are not able to find a uniform definition and explanation of consciousness for us human beings it might be not possible (or a big coincidence) to implement machines with strong artificial consciousness.

2 Overview on Modern Studies of Consciousness and MC

MC attracts attention in different scopes. Starting from philosophy, where consciousness alone is discussed since ages, ending in robotics, where people ask when robots start to be alive. This section brings up the most relevant works in this field and shows how difficult it is to find a common definition of (machine) consciousness. First, intelligence in machines is discussed, without the focus on consciousness. This is followed by modern break-through approaches on consciousness and models of consciousness in machines.

2.1 Assessment of Human-Like Intelligence in Machines

A common definition of *Artificial Intelligence (AI)* is that it can solve tasks, which usually would require some sort of human intelligence. But AI can be seen differently. In this section, there is a short discussion intelligence in machines from the intrinsic perspective. It needs to be questioned if machines give rise to intelligence itself, not only intelligent behavior.

Already nearly seventy years ago Turing discussed about the question if machines can think [Tur50]. For his work, he developed the imitation game or also known as *Turing test*. To limit Turing's work only to the imitation game is a mistake. He additionally mentions building some sort of learning machines and questions how to rebuild the human brain to make it fit into a machine. Important for further cognitive and philosophic considerations are two scenarios. One scenario claims that the brain is a purely deterministic state machine and the other claims that the brain is layered like an onion, even a combination of both is imaginable. Another point Turing mentions is that humans have the ability to scrutinize themselves. This means a human-like machine should also have the ability to criticize received orders due to its own, not-programmed, decision model. This can bring up further discussion since the machine could have its own opinion. Turing already proposed not to develop an artificial brain of an adult human but rather a brain of a human child, which then would have the ability to educate and develop itself. If this would be the objective in MC, there would be two major tasks: (1) Invent the machine child brain and (2) provide the educational process.

A big step forward in the philosophy of artificial intelligence made Nick Bostrom, predicting the date when machines will achieve super-intelligence [Bos98]. As super-intelligence Bostrom defines machines who are more intelligent than humans. Once we have systems like this we would loose control of it very fast. He claims that mankind will achieve such machines in the first third of this century. There are three major dependencies on a super-intelligent system [Bos98]:

1. **Software:** There must be software and algorithms, which allow a super-intelligent system to develop
2. **Hardware:** There must be machinery, which is able to execute the software without big latency.
3. **Knowledge of the brain:** Like Searle, Bostrom claims that before building a super-intelligent machine one needs to know how the human brain is architectured exactly.

2.2 On the Hard and Soft Problem of Consciousness

David Chalmers, a philosopher, also developed a theory of consciousness and addresses problems of detecting it. An important part of his work is the categorization of the problems of consciousness [Cha95]:

1. **The Hard Problem:** The essence of the hard problem is the problem of experience. Every time we think and perceive, the brain does information-processing, but there is also *something* which experiences this processing, a subjective aspect. Already Nagel put this problem under the question “What is it to be a bat?” [Nag74], which only argues that there must be *something* to be a conscious organism. According to Chalmers, this *something* is experience [Cha95]. When we get a sensual stimulus there is not just information processing but also the experience of this processing, even if we are in total darkness we can experience that darkness. This is also known as *P-Consciousness (Phenomenal-Consciousness)*, which is defined by the ability to making experiences [Blo02; Blo90].
2. **The Easy Problem:** These problems are observable through cognitive science. It includes questions like the difference between dreamless sleep and the conscious state or the integration of information by the cognitive system. The easy problem can be categorized under *A-Consciousness (Access-Consciousness)* [Blo02; Blo90].

To put this straight: It might be easy to build a system that does a perfect simulation of a human being and masters the Turing test, but it would be hard to know whether the system has a common experience of its being. Chalmer's quote boils this down:

> *"If any problem qualifies as the problem of consciousness, it is this one. In this central sense of 'consciousness', an organism is conscious if there is something it is like to be that organism, and a mental state is conscious if there is something it is like to be in that state. Sometimes terms such as 'phenomenal consciousness' and 'qualia' are also used here, but I find it more natural to speak of 'conscious experience' or simply 'experience'." [Cha95]* [5]

2.3 Consciousness is Not Provable and Unlikely to Reproduce

Susan Blackmore made it pretty clear what the main problem in machine consciousness is [Bla03; Bla12]. There is no way to test if a machine is conscious. We would require something like the Turing test (used to test intelligence in machines), to prove consciousness in machines. As consciousness, she addresses phenomenal consciousness. It is shown that we are not able to prove consciousness in any organism or object at all. The only thing a human being can be sure about is its own consciousness. Using this experience, the human being implies that his fellows are also conscious. This might be a fact, but there is no scientific way to prove this. Blackmore's work is no proposal for a new theory of consciousness but rather a contribution to show up the problems, which have to be solved before we are able to say that there are real conscious machines. On the top level there are two branches where consciousness can be two possible positions [Bla03]:

1. *Consciousness in a machine is impossible:* This a statements by dualists, materialists and biologists. In dualism, it is believed that there is something god-like, non-physical, which gives us the ability to be a conscious being. Eliminative materialists think that consciousness is not existing and in biology, there are people who think that it needs a fully biological brain to give rise to consciousness.

2. *Consciousness in a machine is possible:* If it is really possible to put consciousness in a machine the steps are pretty clear. First the special ingredient, which gives birth to consciousness, needs to be found and second, it has to be put into a machine.

[5] His theory is explained deeper and broader in his book *The Conscious Mind* [Cha97].

Furthermore, Blackwell evaluates the *Global Workspace Theory (GWT)* [Baa88; Baa97], which is described in section 2.5. She argued that GWT can not explain consciousness [Bla02]. The theory assumes that at any time it can be classified, what is part of the current conscious experience and what not, like consciousness is a container. In her opinion, this would not explain consciousness because in GWT consciousness means to have some sort of illusion, which means that human created machines with human-like consciousness would be a subject of the same kind of illusion [Bla03].

As a result of the evaluation of several theories, she proposed a new theory seeing human beings as memes. Memes are all kinds of behavior that is copied from organism to organism by imitation. For the future, two kinds of artificial meme machines could be considered. Meme machines that imitate humans and meme machines that imitate each other [Bla03].

As a conclusion, it is suggested to solve the problem of machine consciousness by not trying to create machines that already contain human-like consciousness. It would be a better idea to find a memetic co-evolutionary process that will design the machines by themselves, which is similar to Turing's idea [Tur50]. The machines will start to copy information from one to another, maybe exponentially and they will co-evolve with that information. If this would be the case we would not have control over this process or even the outcome of this evolution [Bla03]. Finally, it is seen very skeptical, that we would find the holy grail to consciousness and create conscious machines [Bla12].

2.4 Modern Approaches on Consciousness Studies

O'Regan et al. disbelieve that consciousness is somewhere stored in *Neural Correlates of Consciousness (NCC)*, like Searle [Sea00] and Chalmers [Cha95] suggest or even deeper in a cell's microtubule like Hameroff argued [HP14]. That is why they developed the *sensorimotor contingency theory SMC* [ON01]. Like it already can be assumed by the theory's name, it is argued that conscious experience is dependent on the perception of an outer world. This theory would be in contradiction with Searle's thoughts [Sea90], where it is argued that we cannot be sure that there is no conscious experience in totally paralyzed organisms. Searle showed this in a thought experiment about a patient in a hospital who is in a state physically identical to brain-dead but might still have consciousness, we are just not able to show the evidence of that. Nevertheless there is a work on machine consciousness using the SMC theory. A software model is proposed to implement the theory in robots that are then able to use their own perception and knowledge to generate predictions for their actions

[ME16]. If the SMC theory would be accepted, such robots would only have weak AI (according to Searle) because they would not have intrinsic experience but only observer-relative experience [Sea00]. Related to SMC is an approach by Starzyk et al. viewing consciousness as a metaphysical phenomenom and trying to model human-like cognitive processes and components to make them computable [SP11].

Klink et. al wrote a survey about the scientific study of consciousness, which is a very useful guideline for the research in machine consciousness [Kli+15]. After all, they built the opinion that in the future the neurobiological approaches have the most potential in the research of consciousness since we are facing the age of exponential growth of data, which also holds for neurosciences [Lam10].

The *CRONOS* project is a *"strongly embodied approach"* on machine consciousness [Hol07]. It is pointed out that the main goal in short-term is to increase the performance of robots and as a long-term goal to contribute to the study of P-consciousness. The key features of this model are that the robot builds its own representation of himself *(Internal Agent Model)* and an own representation of the world *(Internal World Model)*. Using this information the robot is able to make its own simulation on a task and later decide, which action is appropriate in the current situation. The resulting robot is functionally similar to a human being. It is also suggested that there could be a chance that the system they provide already contains a glimpse of P-consciousness [Hol07].

Because there are some overlaps with IIT [TK15] panpsychism is proposed as another theory of consciousness. It asserts that every real object contains an own mind (not to be confused by an own brain) or a mind-like quality (e.g. a stone can have some sort of experience as well as our planet). Panpsychism suggests that everything in the universe has an intrinsic existence, which is seen as consciousness. In this scope it is very important to see conscious objects not as organisms having feelings and thoughts, it has to be seen as something much more abstract and subtle. There is a very detailed work, which argues for taking panpsychism serious with the help of works from the last four centuries [Skr03].

2.5 Global Workspace Theory of Consciousness and LIDA

The *Global Workspace Theory* (GWT) was introduced by Baars and is settled in the field of cognitive science [Baa88; Baa97]. The motivation for developing a new framework to describe cognitive consciousness was the need for an empirically testable theory of consciousness. Philosophical models might be right or wrong, but in cognitive sciences, it is important to make testable experiments. The philosophical models are often too complex and not computable [BF07; BF09].

The *Global Access Hypothesis* is the main assumption of the GWT. It claims that conscious contents cause widespread brain activation. The second assumption is the *Working Memory (WM) Hypothesis*, which defines that conscious contents cause unconscious WM functions, which are needed for verbal, visual, and executive functions. Like mentioned before, Blackmore had a counter argument to this hypothesis (ref. section 2.3). The goal of GWT is to specify the role of certain brain activities in cognition [Baa88; Baa97].

There is a suitable metaphor to describe the basic features of GWT (ref. figure 1): In a classical theater, there is a stage and the audience. The people in the audience are unconscious processors. The stage builds the WM containing a conscious spotlight that broadcasts to the audience. The attention of the organism controls the spotlight on the stage. In the backstage, there is the operation of all unconscious states that contribute to shape and direct the conscious contents [Baa88; Baa02].

The *LIDA model (Learning Intelligent Distributed Agent)* is an implementation of GWT, also known as *Intelligent Distributed Agent (IDA)* [Fra+12]. The idea was to implement an autonomous agent in a given environment. The agent would sense his own model of the outer world and act to it. The agent's actions must also influence, what it will sense in the future, therefore the agent is bound to the environment. It is believed that future insights in the cognitive studies of consciousness are sufficient to create autonomous artificial agents. Originally IDA was developed for the US Navy to perform tasks that would usually require trained experts [FP06]. GWT and LIDA claim that conscious cognition is necessary for complex learning in organisms and agents. The core of IDA implementing GWT is a cognitive cycle, which can be divided into nine steps [FKM98; Fra00; FP06; Fra+12]:

1. A new sensory stimulus is received and filtered to add meaning and produce a perception.
2. The new perception is sent to the preconscious working memory, where already past perceptions are stored (which will decay over time) and a high-level perception is built.
3. The structure working memory triggers temporary episodic memory and declarative memory, which produces local association that ist stored in the long-term working memory.

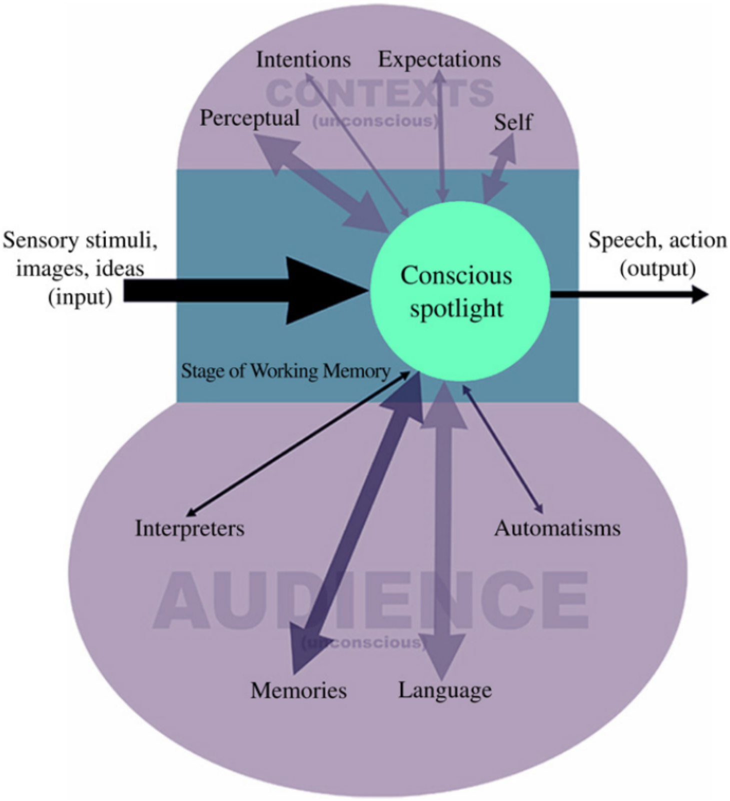


Figure 1: Architecture of the GWT using the theatre metaphor. Consciousness is a moving spotlight on the stage of the WM. The backstage builds the context, which is sub-conscious and there is broadcasted information flow between the stage and the audience containing memories, language, automatisms and interpreters [BF07].

4. A cluster of contents in the long-term WM competes for the spotlight of consciousness, which means that the most relevant or urgent task receives most of the system resources.

5. According to GWT, a conscious broadcast is sent to enable different forms of learning as well as the reservation of internal resources.

6. After the broadcast, the procedural memory is responding to it.

7. Other unconscious components are responding and instantiate copies of themselves to select the appropriate action, do the variable bindings and the activation passing.

8. The correct action for the current cognitive cycle is selected.

9. IDA executes the selected action to the external or/and internal environment.

After all the abstract theory about GWT and LIDA it is refreshing to tell that there is already an application to a real problem using GWT [PRG16]. A machine consciousness approach is used for urban traffic control. GWT is implemented to optimize the circuit for traffic lights.

There is one big difference between GWT and IIT to mention. IIT addresses P-Consciousness while GWT addresses A-Consciousness. This means that IIT is a theory to find out why experience is in certain systems and how to give rise to it, while GWT wants to replicate human behavior to study cognitive consciousness. Nevertheless, the theory might be useful to develop machines with simulated consciousness.

2.6 Further Readings on Machine Consciousness

There are three recent works, which focus on assessing different theories on machine consciousness written by Gamez (2008), Reggia (2013) and Sun et al. (2011) [Gam08; Reg13; GS12]. The next paragraphs summarize the core statements out of these works. Additionally, there is a work by Chella et al., which builds a bridge between studies on AI and consciousness [CM07]. A key message here is that it would be difficult to predict the consequences if we would be able to create a machine with human-like consciousness (no matter what the real cause for consciousness is).

An extensive work about MC was written by Gamez. He showed the relevance of research in MC and the challenge of the highly interdisciplinary research area, which includes philosophy, psychology, physics, neuroscience and computer science. Besides the presentation of a qualitative selection of the works in this field (including some approaches presented in this work) he presented four different levels of studying MC [Gam08]:

1. **MC1**: Machines that only behave like they would be conscious.
2. **MC2**: Machines containing cognitive characteristics that are similar to consciousness.
3. **MC3**: Machines with the same architecture, which is correlated to the rise of consciousness in organisms.
4. **MC4**: The study of machines with P-Consciousness.

It is shown that MC2 and MC3 can already contribute to more intelligent machines, but there is no evidence of machines supporting MC4.

The conclusion of another wide-range review by Reggia [Reg13] is that computational modeling is an accepted and effective tool to study different theories of consciousness. Also, computational models could confirm a number of correlates to consciousness in neurobiological, cognitive and behavioral studies, but the author thinks that this will not lead to machines having real P-Consciousness. Furthermore, it is concluded that currently there is no existing approach on real artificial consciousness in a machine or even a valid forecast that consciousness in machines would even be possible, which is similar to the conclusion by Blackmore (ref. section 2.3). None of the current studies, even if claimed otherwise, would have a satisfactory argument to explain how the approach to real machine consciousness should be studied or how studies could lead to artificial consciousness.

There is also a philosophical assessment of computational models of consciousness addressing six different approaches: *Clarion*, *LIDA*, *ACT-R*, *NWS*, *ART* and *GMU-BICA* [GS12]. According to this study, *LIDA* implements most of the relevant aspects to be able to provide conscious experience (according to philosophy), but still, there are many improvements to make. Anyway, it is highly promoted that computational modeling theories of consciousness will accelerate new insights in the study of consciousness, while the final implementation of real artificial consciousness would initiate the end of conscious studies. It has to be mentioned that this work only addresses the philosophical side of the selected works and did not implement the model itself.

This chapter introduced diverse theories on consciousness and machine consciousness. There can also be seen parallels between different approaches and theories. People like Searle, Chalmers and Tononi see consciousness as the ability to have an real experience [Sea00; Ton04; Cha95]. Despite there are being some good approaches to model consciousness in machines, there is no evidence about works, which prove that there is already a machine with an own experience (P-Consciousness).

3 Background on Consciousness Studies

This work deals with three scopes around MC to be able to find a properly modeled implementation of simulated conscious machines. (1) Since it is not possible to prove a scientific definition of consciousness, a strong philosophical basement must be given. (2) There needs to be a theory of consciousness, wich is applicable to artificial systems in computer science. (3) Finally, the theory has to be integrated into a simulation framework, where it is possible to observe simulated consciousness. In this section, a brief introduction is given to cover such three points.

3.1 The Maze of Consciousness is Solvable

To build the philosophical basement it is referred to John Searle. John Searle is a modern philosopher who not only develops theories on consciousness but also forms ideas about conscious machines. The big difference between consciousness and other phenomena in the biological or natural world are the three aspects of quality, subjectivity and unity [Sea00]:

1. **Quality:** It is a different between sitting in a room looking on a candle or sitting on the beach observing the sunset.
2. **Subjectivity:** Consciousness only can arise in a subject. Without a subject, there is no experience and therefore no quality of consciousness.
3. **Unity:** This follows from the fact that quality and subjectivity only can co-exist and not be seen independently.

First, it should be discussed where to place modern AI. In an early work, Searle argued to differentiate between strong AI and weak AI [Sea90] (ref. section 3.1). In weak AI computer systems are not having real intelligence, a real mind or some sort of sentience. Machines with strong AI would be systems with an own mind and all the features a weak AI do not have. Furthermore, it is shown that if there would be strong AI it not just might have intrinsic experience but also must have its own feelings and thoughts.

A computer program is operating syntactically, while the mind operates semantically. Therefore software is purely formal, constructed by a sequence of symbols (e.g. a special order of zeros and ones), which is also known by the syntax of the software. The meaning is only added by the observer, in this case the observer is the computer who interprets the syntax. On the other side, human minds have mental contents, every state of mind has a certain meaning (semantics). Searle argued that syntax is not able to generate any semantics, hence software can not generate human minds. To address the importance of the difference between syntax and semantics. Searle shows that syntax alone is not sufficient to generate semantic content. If we generate a software, which can only be out of formal syntactic symbols, the observed consciousness is not intrinsic to physic but rather observer-relative. For the creator, it has a meaning but for physics, the software only generates different electronic signals. This should show that computational models of consciousness alone cannot be conscious by themselves, independent of their physical substrate [Sea90].

Another argument related to the *syntax-semantics-problem*, Searle developed, is that brains cause minds, therefore also artificial conscious systems must have its own causal powers. It would not be enough to start a computer program called *consciousness* to blow a soul into a machine. The whole machine might be architectured in a way that consciousness can evolve from its physical structure. Hence, also strong AI has to have causal capacities, like human brains they would cause a mind and not be the mind.

> *"Is the Brain's Mind a Computer Program? No. A program merely manipulates symbols, whereas a brain attaches meaning to them." [Sea90]*

From the first point of view, the above quote might sound sad for people who work on machines with P-Consciousness, but it only refers to computer programs as not part of the mind like mentioned before. It is better to see this statement as a motivation to think about how it would be possible to provide the hardware, which could give rise to artificial consciousness. Unfortunately, there is no evidence that we are even on the starting line to solve this mammoth task.

For people, it is hard to understand that the mind is only a biological phenomenon. They fail to realize that it is in the same area as the phenomenon of digestion or growth [Sea90]. That is also the reason why people might think that a simulated consciousness might be a real consciousness. It is much simpler to believe in some sort of dualism, that there is some special ingredient which is not scientifically detectable and is the cause of consciousness, as to believe that the mind is only the result of very difficult neurobiological processes. That is why simulated minds are often mistaken to be real minds.

As far as we know, everything which is part of our experience is based on cause and effect. This experience is reducible on how the neurons in our brain are firing, how they are connected to each other and in which order they are firing. This is somehow the cause of all conscious life we can observe in our external world. All stimuli of lifeforms are converted by our nervous system into a system of firing neurons [Sea90].

Based on the belief that the Turing test can lead to strong AI people mistakingly might think that consciousness could be analyzed *behavioristically* or even *computationally* [Sea93; Sea00]. Therefore it would be sufficient to have the right software transforming given inputs to the right outputs. It is clear that conscious states usually cause behavior. But it has to be distinguished from the conscious state of the mind and its effect [Sea90]. The *Chinese Room* thought experiment can be used to explain the difference: Imagine an isolated room where someone is sitting with a complete dictionary of Chinese phrases. There are two slots to the outer world. One where the person receives a question, written in Chinese and one where the people send the answer, also written in Chinese. The person itself is a human being only able to speak English. But the Chinese phrase book makes him able to answer all the questions. From the outside perspective, it would be like there is only a computer in that room who understands the question and gives the proper answer, therefore the machine would pass the Turing test. But there would be two contradictions: (1) There is a conscious being in that room and (2) that person would not have any understanding of what the input (cause) and output (effect) is. This example should show that consciousness should not be seen behaviouristically or computationally.

To stay realistic, Searle mentioned, like Susan Blackmore (ref. section 2.3), that the only conscious state we can observe is our own [Sea97]. Everything else are only implications. We imply that our fellows also have conscious states because they act similar to us. Another thought experiment should undergird this argument [Sea97]:

Let us assume somewhere in the future mankind is able to create robots having their own real consciousness. Now people order some robots for a special task in a factory. It turned out that the created robots were not fulfilled by working on the task and they turned to be unhappy. To redeem the conscious robots, they create other robots without consciousness having completely the same behavior as the conscious robots but would not have any sort of experience, nor feelings. In that case, the behavior was separated from consciousness, but only the creator of such machines could identify the robots with consciousness.

In the last work by Searle, discussed in this thesis, he made a prediction where consciousness could be found in physical lifeforms. He also tried to give hints for further research how to make a breakthrough in conscious studies

[Sea00]. Since conscious states have their cause in neurobiological processes they are a high-level feature of the brain. This means the big task will be to solve the question of how the brain generates the cause to conscious states and how the brain is able to realize that itself is in a conscious state. Searle (2000) mentions *Neural Correlates of Consciousness (NCC)* as an research area with serious potential to contribute to conscious studies [Sea00]. The research of NCC would pursue tow separate problems:

1. Find the neural correlates of consciousness.
2. Find the cause of consciousness in NCC.

It is possible to address these questions at different levels. At the level of single neurons or synapses, or on the level of communities of neuronal maps or even on the level of whole clouds of neurons [Sea00]. Maybe also a combination of all three of them, like Turing suggested by imagining that the system in the brain could be layered [Tur50]. It might also be possible that the level is still too high and it has to be done more research in deeper levels as the level of neurons and synapses, e.g. on the metaphysical level [HP14]. Even tough there are a lot of open questions Searle was sure that *"...Consciousness consists of inner, qualitative, subjective states and processes of sentience or awareness..."* [Sea00]. Finally, there are three steps to investigate consciousness if neurobiology is seen as the key to understand consciousness [Sea00]:

1. Find the NCC of consciousness.
2. Simulation and testing of the correlates for their causal relation.
3. Formalize the causal relationships in the form of a theory.

3.2 The Integrated Information Theory Makes Consciousness Testable

For this thesis, IIT was selected for the modeling and implementation of collective machine consciousness. While this theory is relatively young for studies in consciousness, the first publication appeared in 2004 [Ton04], it already has two revisions [Ton12; OAT14]. Before describing the core aspects of IIT, it is explained why this theory is of a keen interest in consciousness studies and why it was the choice for this work. To understand the essence of IIT one should know its roots. The theory was developed in cognitive studies, led by Tononi [Ton04]. His studies are focused on neuroscience and sleep research

“Consciousness is everything we experience.” *[Ton04]*	“...differences that make a difference...” *[Ton12]*

Table 1: In IIT consciousness is defined by axioms of consciousness and their postulates for physical organisms. Anyway Consciousness should be seen as pure phenomenal experience as stated in the left quote. As defined in the right quote experiences can be distinguished by their differences from other experiences.

where he observes brain activities in conscious and non-conscious states of beings.

Most researchers are driven by the challenge to find a scientific definition of consciousness, but Tononi et al. were driven by the challenge to make consciousness testable, which is a similar motivation like in the GWT. This is relevant in many clinical cases, e.g. to test if a coma patient is really brain dead or if the patient has still some degree of consciousness. As their work was interpreted for this thesis, it was less important to make a clear zoned definition of consciousness (ref. table 1), but rather to find clear rules (axioms) about consciousness.

3.2.1 Information and Integration Explained

In IIT for a system, it is necessary to integrate information to have some sort of experience. This section deals with the explanation of how to understand integration and information according to IIT. All these aspects of experiences are best described by metaphors inspired by [Ton04].

Information: Imagine there is a new sensor on the market who can distinguish between coffee and tea. If you dip the sensor into coffee it returns 0 and if it is tea it returns 1. Of course, if you would drink the liquid it is easy to tell if it is coffee or tea. Despite the input and result is the same, comparing you testing the liquid and the sensor testing the liquid, there is a huge difference in the information processing. The sensor might be built to search for the one special molecule contained in coffee beans and leave out all the rest of the information. But you are processing way more information while tasting coffee, you can judge if you like the taste of the coffee or not and might have a memory of drinking coffee last summer in Italy on the beach or might have a link from drinking coffee during long study nights. The difference could be measured by entropy and because we would have an immense higher cause-effect-repertoire compared to the sensor, which can only return zero or one, the entropy of a human being would be exponentially higher. It is an interesting fact, that viewing information as experience and as an important part of consciousness has yet not been the focus of conscious study [Ton04].

Integration: To extend the above thought experiment a grid of $100 \times 100 = 10,000$ different sensors each measuring a different molecule is imagined. This means, assuming binary sensors, there would be $2^{10,000}$ different states and as many different liquids, the sensor cluster could theoretically detect. In the background, there could be a lookup table to determine if the liquid is coffee, coffee with milk, green mountain tea, Darjeeling tea etcetera. The difference between the sensors and human experience would be that we would always taste the whole aroma. Maybe you are able to distinguish if there are milk and sugar added but you can not separate the coffee taste from the milk and sugar taste. The sensed liquid is experienced in a whole, as integrated information. There are dependencies between all the different tasted stimuli, which means that some elements are causally dependent on others. On the other side, the sensor cluster is not integrated and would not have a full experience. It would simply aggregate the sensed and not-sensed molecules.

To put that together, according to IIT a conscious system would have the ability to experience. Experience can arise if the system processes information and is able to integrate that information. This would also mean that information is able to change the state of the system (*"... differences that make a difference ..."* [Ton12]).

3.2.2 Postulates and Axioms of Consciousness

Like mentioned before IIT is built on postulates and axioms of consciousness. There are often cases where postulates can be the same as axioms but in IIT they have different meanings. With the publication of IIT 3.0 postulates and axioms are introduced [OAT14]. Axioms are undeniable facts about consciousness, while postulates are an assumption about the physical world. In IIT postulates define the requirements for physical systems to be able to give rise to consciousness. In the following table (ref. table 2 on page 22) there is a brief introduction to the five different axioms and postulates of IIT. It is possible to differ between postulates of mechanisms and postulates of systems of mechanisms but for simplicity, only the mechanisms' postulates are explained [OAT14].

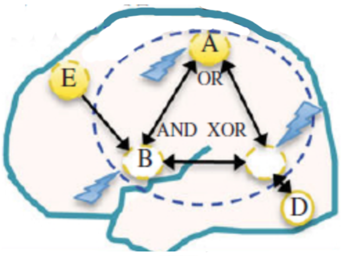
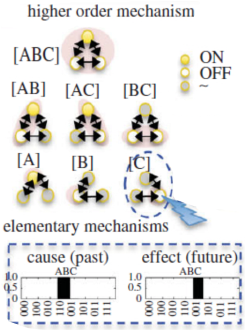
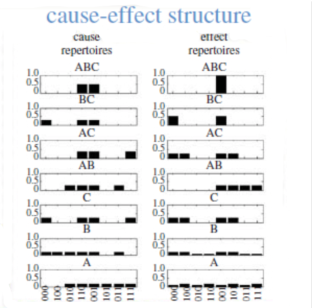
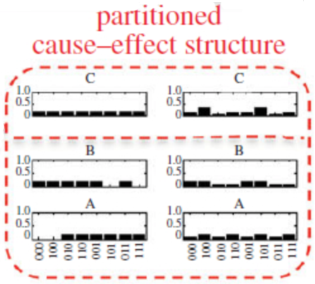
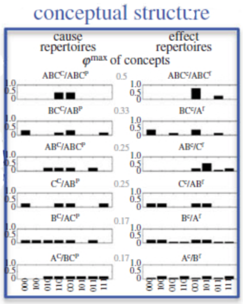
	Axioms ⇒	Postulates	
	Existence		
	The only thing about consciousness we can be sure about is that it exists. Inspired by Descartes' *I know that I am, since I can make experiences of the outer world.*	A system contains a set of mechanisms, e.g. neurons in the brain. A state is formed by the mechanisms of the system.	
	Composition		
	Every experience is a composition of multiple objects. An experience is made in a room and the room could contain a chair, a table and other people. It would be a different experience if we would change only one single object.	Low-level mechanisms can be combined into high-level mechanisms, e.g. a XOR mechanism can be built by AND and OR gates.	 higher order mechanism [ABC] ON OFF [AB] [AC] [BC] [A] [B] [C] elementary mechanisms cause (past) effect (future) ABC 1.0 0.5 0
	Information		
	Conscious states give information, experiences differs from other experiences by different information, which means that also complete darkness has information and can be experienced.	A mechanism only contributes to consciousness if it constrains the states of the system, the mechanisms should make a difference to the system.	
	Integration		
	Experiences can not be divided and contain irreducible components. If you see a green car driving by, the experience can not be divided into only seeing the car or the color or even just experience the speed of the car. It has to be seen as a whole.	A mechanism only can contribute to consciousness if it it is integrated into the system, e.g. if deleting the mechanism would not change the experience, the mechanism is reducible.	
	Exclusion		
	Experience has a spatial and temporal attribute. We only can experience certain things at certain times. For example We can not experience how it is to be on the moon or can not replay our birthday.	A system only creates a unique cause-effect repertoire, which also has the highest integration φ^{Max}.	

Table 2: Axioms and Postulates. The aximos about consciousness help to find requirements (postulates) for the the physical system to give rise to experience. In IIT there are five axioms defined: Existence, Composition, Information, Integration and Exclusion [TK15].

3.2.3 Measuring Integrated Information

Consciousness according to IIT is measurable. The problem of computing the quality and quantity of consciousness in IIT has to be split up. First, the network has to be taken into account, which consists of nodes and edges. A network is the model of an organism's brain (e.g. a network of neurons). Second states of nodes have to be considered. A state is a specific value for each node at time t_i, which causes the next state at t_{i+1}. The base of the calculations is the transition probability distribution of all possible states in the system. For a given state at t_i the probability for every other state at $t_i + 1$ is given. Using this input the *cause-information (ci)* as well as the *effect-information (ei)* can be calculated for a mechanism (or a set of mechanisms). Further calculations lead to the *cause-effect-information (cei)*, which is the amount of information stored in a mechanism. The distance between the probability distributions is calculated by the so-called *earth-mover's distance (EMD)* [RTG00].

Integrated information is essential for an experience. Without information that differs from other information, no new experience can arise. There are two types of measures, which can be assessed: *Integrated Information* (φ) and *Integrated Conceptual Information* (Φ). φ measures how much a single mechanism (e.g. neuron or logic gate) is integrated into a system by measuring its irreducibility. On the other side, Φ measures the integration of a whole concept, a set of mechanisms. Of interest here is the investigation of a Main Complex (MC), which is a set of mechanisms with the maximum integrated information Φ^{Max}.

For a mechanism to have integrated information (φ) it is necessary that it has integrated causes as well as integrated effects. The mechanism would be not integrated if there is a mechanism in the system, which is not affected by the mechanism. This is explained by an example system in figure 2 on page 24. Iterating over all possible purviews, the highest possible integrated information is defined by φ^{Max}, which is the quantity for the amount of integration (or irreducibilty) of an mechanism. A purview contains each cause and effect repertoire of each possible set, contained in the power set of the system, at a certain state for a given node. In figure 3 on page 25 it is shown how Φ can arise in systems of mechanisms. For a system with $\Phi > 0$ it is necessary that all mechanisms are strongly connected to each other.

There is much more to learn about the calculations of the IIT, but it would definitely go beyond the scope of this work. In the later evaluation of the experiments, it is focused on values of $\sum \varphi^{Max}$ and Φ^{Max}, which were introduced in this section. It is referred to IIT 3.0 for further information about the computations [OAT14].

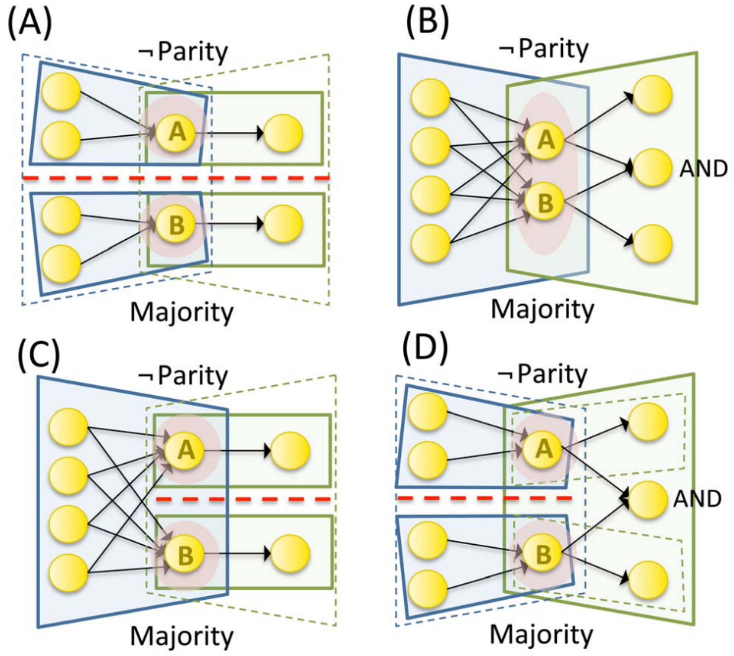


Figure 2: Looking at the joint mechanism AB, it is not existing from the intrinsic perspective in (A) since both have independent causes and effects. In (B) AB are are integrated and therefore generated integrated information, which means that the joint mechanism AB exists intrinsically. In (C-D) AB might have integration in the past or in the future but exists not intrinsically [OAT14].

3.2.4 Supports and Critics of IIT

There is already a vibrant discussion on IIT. Besides neuroscientific and philosophic supports, there are also critics. The following paragraphs show a selection of this discussion, but will not assess the diverse opinions.

Cerullo published the most detailed critique about the integrated information theory [Cer15]. His main critic is that there is no evidence for Tononi's claim that information exclusion (ref. last row in table 6) is self-evident. Furthermore, in his eyes, it is also a problem that intuitively non-conscious systems would have consciousness, which points to the panpsychism attributes of IIT. A final statement of Cerullo is that researcher of AI and consciousness can be sure that "... IIT does not banish the ghosts from their machines." [Cer15].

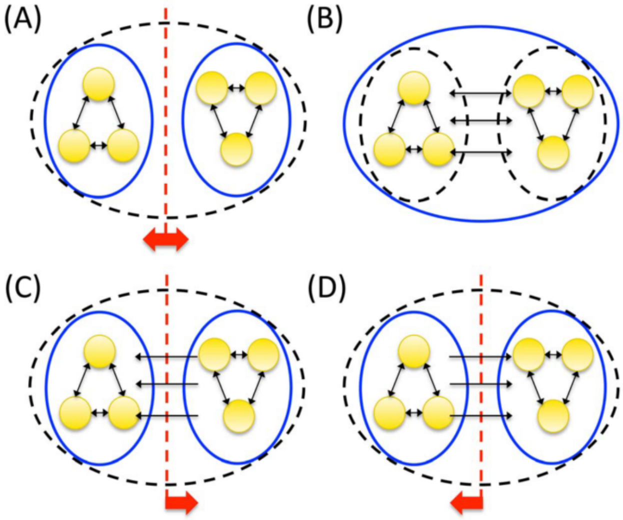


Figure 3: For integrated conceptual information Φ it is necessary that all elements in a set of elements have causes and effects in all other elements. The black dashed line shows the border of the system. (A) The system can not be integrated since the two systems of mechanisms are not connected. (B) The system could have Φ since the systems of mechanisms are integrated and a state change of one mechanism causes a change in all others. (C-D) Only one subsystem is affected by state changes of the other subsystem, which means that no integration of the whole system is possible [OAT14].

Maguire et. al question the computability of integrated information [Mag+14] similar like Turing (1950) questioned if computers are able to think. It is discussed that conscious states and therefore integrated information can be described by data compression. Using known theorems about data compression proves are given that integrated information is not computable, in other words, current machines could not have integrated information [PRP16; Mag+14].

Gilipithis develops a novel theory of consciousness called the theory of Noémona species, which can address not only human consciousness but also possible animal and machine consciousness. In the latest work it is claimed that Tononi's theory is vague and there is a huge gap between the key aspects of IIT that (a) experience is integrated and (b) experience is information. It is accepted that the key aspects are true but that the mathematics of IIT can not bridge the gap. There is no further explanation for the reason of this claim [Gel14; Gel09].

Max Tegmark, a physician, developed a theory where consciousness can emerge by rearranging matter, from the physics perspective a human being would be only rearranged food [Teg15]. It is explained that no extra ingredient is needed to add to physics, as we know it, to give rise to consciousness (nondualistic view). This would mean that not the particles itself matter but the pattern how the particles are arranged matters [Teg15]. This theory refers to IIT and views integration of information also as a key principle for physical conscious systems, hence it can be seen as an extension of IIT in the field of physics.

To bring together the rather abstract IIT works written by or contributed by Christoph Koch can be reviewed. He focusses on the neurophysiological investigation of consciousness and the NCC. Recently it could be shown that there are some hot zones in certain parts of the brain, where consciousness could arise, but there is no unique sector, which is responsible singularly for conscious states [KT07; Koc+16; Cri+04].

3.2.5 Animats with Integrated Information

Animats are *artificial agents with intelligent behavior*. They are used in this work to evolve and evaluate agents with integrated information. Animats usually have three types of units [MHA13]: (1) **Sensors** receive information from the environment. (2) **Motors** make effects on the environment and change the extrinsic world. (3) **Hidden units**, who play the role of a brain and memory.

In works by Albantakis et al. [Alb+14] as well as Edlund et al. [Edl+11] it is shown that it is possible to evolve artificial animats with simulated consciousness by using *Genetic Algorithms (GAs)*. Additionally, it was investigated, that it is more likely for animats with high fitness to have integrated information as for not integrated animats. Albantakis et al. use a simulation model with varying task complexity and varying architecture of animats. It could also be shown that on tasks, which require more memory higher integrated information is developed. Additionally, it is possible to observe, that constraining the amount of sensors, forces the evolutionary process to create systems with higher integrated information to be able to achieve high fitness [Alb+14].

3.2.6 Alternatives to Calculate Integrated Information

The proposed formulas in IIT 3.0 are not the only way to calculate integrated information. In the last few years there came up a set of alternative computation models, with the intension to be more practical and applicable to real data:

- An approach, which makes it possible to calculate integrated information for neural systems and time-series data gathered by neuroimaging [BS11].

- Oizumi et al. introduced an additional measure $\Phi*$, which measures the amount of information lost during the partitioning of the system [Oiz+16]. $\Phi*$ is already applied successfully to classify visual experiences by using information of direct human brain recordings [Hau+16].
- Krohn et al. try to make a general formulation for the computation of integrated information. They use probabilistic models to implement their approach [KO16].

3.3 Collective Behavior Through Evolutionary Algorithms

To implement artificial animats with collective behavior *evolutionary algorithms (EAs)* are used. The system optimizes an organism towards one or more fitness functions. The organism contains several attributes, which changes the behavior and therefore the fitness itself. A special case of EAs are GAs. In a GA an organism is constructed by its genome and in the optimization procedure, the genome is mutated, not the single attributes itself. In figure 4 the basic workflow of an EA is explained. EAs are metaheuristics. Such algorithms might not find an optimal solution for a given problem but try to converge to it. They are applied to problems with a huge search space, where it would take a disproportionate amount of time to find the optimal solution.

The collective behavior usually arises in organisms of the same species. From the outside perspective, it can be seen as one collective movement of a group of individuals, while there is no central unit controlling the swarm. Each organism moves according to simple rules and therefore contributes to the collective behavior. This can lead to a complex and intelligent behavior. Collective behavior and swarm intelligence have its roots in biology. A swarm of flying birds seems to be a highly complicated construct, but in reality, they only act to local rules [GGT07].

It is common to use EAs or GAs to develop multi-agent systems, which have the goal to perform a collective behavior or even show swarm intelligence [KMS09; Mii+12; Ols+12; Ols15]. Besides defining the right model of the agent, animats in this thesis' terminology, it is important to find the correct fitness function, which leads to the evolution of collective behavior. Usually, it is better to not benefit the detailed behavior of the agent but benefit the agent if it contributes to an overall goal [WT11; Lim+96].

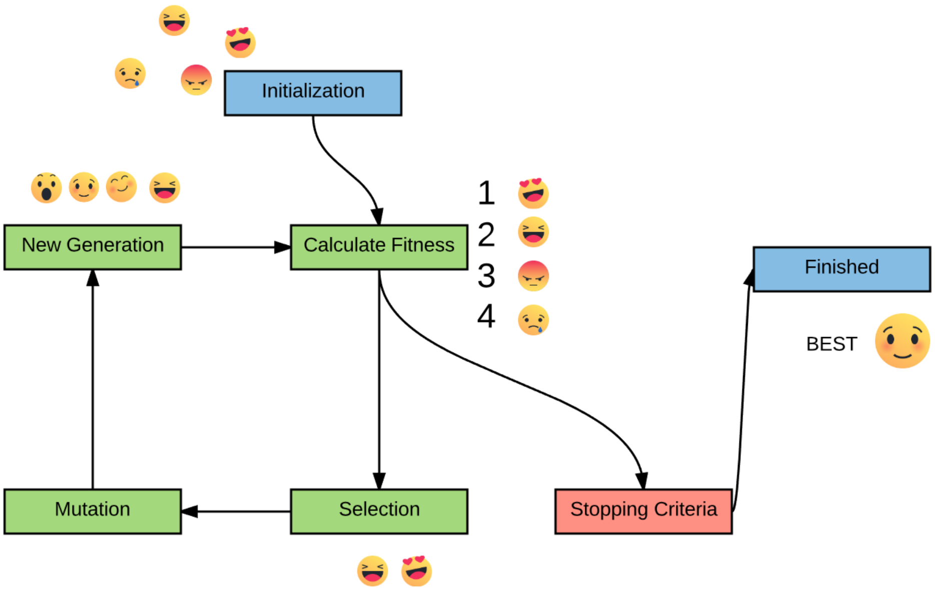


Figure 4: Basic workflow of a EA. (1) The algorithm is initialized with a random population. (2) The fitness of each population is calculated. (3) Select good performing organisms for optimization (4-5) Generate a new population by mutating the selected organisms. (6-7) If a stopping criteria is met the algorithm returns the best found organism.

3.4 Markov Brains as Finite State-Machines

The processor unit of an animat, its brain, can be of different types. The brain could be simply a manually coded function, an artificial neural network or, like in this work, Markov brains [Edl+11]. The major constraint for animats is that they need systems that can transform the given input to a resulting cause, but due to the special calculation models in IIT, further constraints about the brain are given.

According to IIT conscious systems have to major constraints. First, it is necessary that the mechanisms in the systems are integrated and second the system must be state-dependent. In Markov brains, it is possible to implement integrated systems, which means that each mechanism could influence each other. Additionally the mechanisms in Markov brains can have states, which is dependent on the past input. This simply means that it is able to create a memory in the brain. This two features, integration and state-dependency, are implemented with the support of *Hidden Markov Gates (HMG)*.

In figure 5 on page 29 there is a sample architecture of a HMG shown. The input of a gate is given by a set of input nodes. Each node can have a state (e.g. 1 or 0). Inside the HMG there could be a lookup table, or any other mechanisms, transforming the input into an output (a world feedback). The output is written to a set of nodes and form the input for the next time step. Figure 6 on page 29 shows a sample architecture of a Markov brain. The state values for each node at t_i is sent to connected HMGs. After interpreting the inputs, each HMG sends the new state to the same set of nodes at time t_{i+1}.

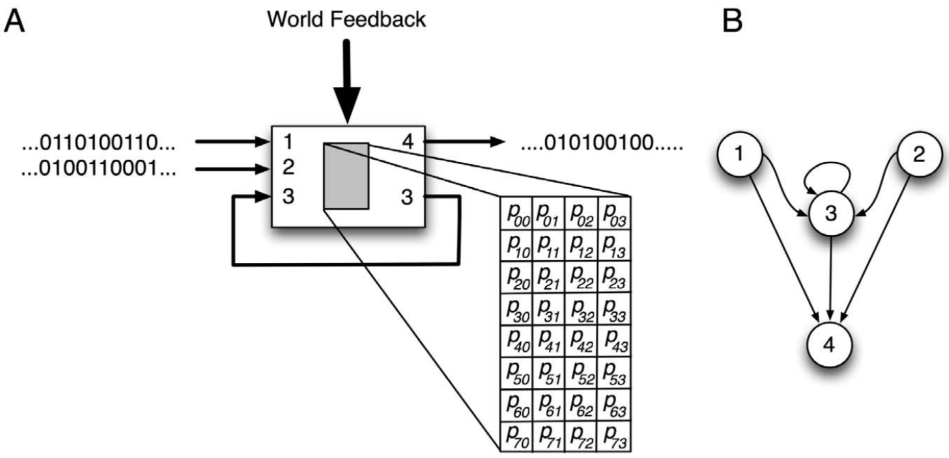


Figure 5: Architecture of a HMG. (A) shows a HMG, which has a set of inputs and outputs. The inputs are received by connected nodes at t_i and the outputs are sent to nodes at t_{i+1}. A usual gate contains a deterministic or probabilistic lookup table, which maps the input state to an output state. (B) Depending on the node connections, the HMGs construct the network structure of the Markov brains [Edl+11].

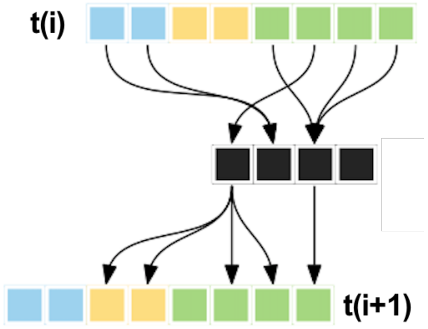


Figure 6: Architecture of a Markov brain. Blue boxes show the input nodes. Yellow Boxes show the output nodes. Green boxes show the hidden nodes, where knowledge can be stored. The states of the nodes at t_i are sent to the HMGs. The gates transform the received cause to an effect and send a new state to the gates at t_{i+1}.

3.5 Can a Machine Develop Real Consciousness?

A machine could develop real artificial consciousness, if we find a new physical architecture for machines and if dualists are on the wrong paths. In IIT consciousness is defined as experience and a system of integrated mechanisms can give rise to consciousness [Ton04].

As it was argued before, it is not sufficient to only develop the right algorithms and software. For real phenomenal consciousness, the system must exist physically. In comparison: For humans, simulated systems might be totally intelligent and seem conscious (observer-relative) but intrinsically and physically the system is nothing else but a very powerful calculator [TK14; Ton+16; TK15; Sea00].

> *"Conversely, digital computers running complex programs based on a von Neumann architecture would not be conscious, even though they may perform highly intelligent functions and simulate human cognition." [Ton+16]*

Robots have to be reinvented. Currently, the CPU, memory and working memory in computers are not integrated, which makes it easy to replace them without changing the structure of the system. Future artificial conscious systems might combine all these components in an integrated manner, hence they might be comparable to biologic organisms.

It is also necessary that mankind accepts to discuss different forms of consciousness. Talking about machine consciousness, people dream big and imagine machines with human-like experiences, the abilities to have fun or machines which are creative or develop love. But it would be really limited to believe that only mankind would have consciousness, it also has to be worked on consciousness in animals and plants or other organisms. Having more knowledge about the consciousness of real world organisms, it might be possible to find one architecture simple enough to clone it into an artificial machine.

Finally, it is not provable, that it is impossible to create an artificial machine. But with a look at the current research going on, we are still at the starting line and not far ahead of (or even on the same level) Turing (1950). Still, there is no provable definition of consciousness, since any theory is only disprovable. Despite IIT has promising contribution into that direction, there is a need to develop approaches of machines where we can implement real artificial consciousness.

4 Model to Evolve Conscious Animats with Collective Behavior

In the second part of this work, IIT should be investigated in simulated collective behavior. Therefore a simple task was modeled, which should encourage collective behavior. The model consists of three different tasks:

1. Design the animats' environment.
2. Design the animats acting in the environment.
3. Design the optimization function, which allows evolving collective behavior.

Furthermore, it should be possible to investigate integrated information in the brain of the animats. To make statements about influence factors of the development of integrated information, there should be animats varying in their sensory- and motoric-abilities. There should be variations in the group size of the collective to make further investigations, too.

4.1 A 2D-Matrix Environment

The challenge in designing the world, animats are acting in, was to find an environment where multiple animats are able to co-exist and evolve collective behavior. Therefore four different constraints were defined to find a proper environmental design.

1. Animats must be able to co-exist.
2. Cooperation, non-egoistic behavior, helps to gain higher fitness.
3. The task is complicated enough to evolve integrated information.
4. The task is simple enough to solve it with only lightweight animats (having just a low count of sensors, effectors and hidden units).

Using this restrictions and rules, to design the environment, various different configurations were tested. First, the idea was to develop a real world simulation of army ants building a living bridge (e.g. [CF03; Rei+15]). It turned out that the task was too complicated and animats could not be developed in a simple manner. That is why the decision was made to develop a world inspired by Koenig et al. [KMS09]. In a limited room of 32×32 blocks, animats are placed in predefined start slots (ref. figure 7). In the middle of the room, there is a gate, which divides the room in two similar sized sub-rooms. If an animat steps through the gate, it receives a reward. If an animate collides with another one it gets a penalty, but only the animat who made the movement.

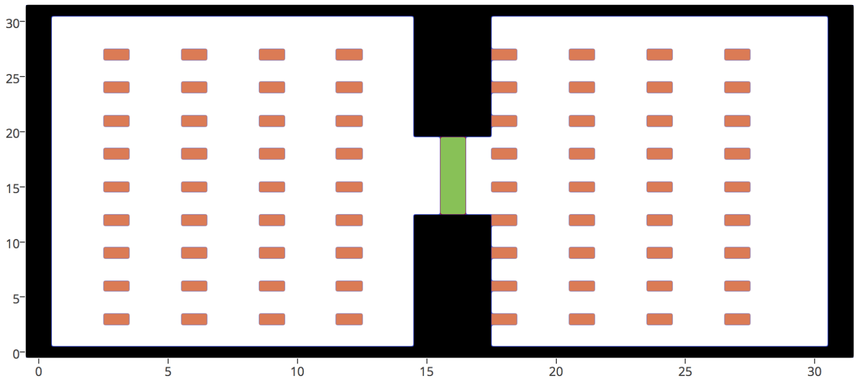


Figure 7: The level is designed as a 32×32 matrix. The black cells are boundaries, which are not crossable, the red cells are start slots and the green cells mark the goal.

4.2 Six Different Animat Designs

Since the goal is to investigate possible causes for the development of integrated information it is important to design a set of unique animats, where the animats differ in a particular way from the others. For this purpose, we came up with six different animat designs. The animats were identified after greek letters from α to ζ. It should be distinguished how much the count of hidden nodes and the count and type of sensors and motors would influence the development of integrated information.

It is obvious, that an agent with a large number of sensors and hidden units can archive a higher performance as an animat with only one sensor and limited hidden units. But the focus of this work is on the investigation of integrated information in collectives and not on the development of optimal performance. Therefore it is assumed that the measures of integrated information will vary, while the group performance stays on the same level comparing different architectural animat models.

In table 3 the architectures of the distinct animats are listed. Green symbols mark sensors, red triangles mark motors and yellow circles mark hidden units. Three different brain capacities (2, 3 and 4 hidden units) and two different sensor settings are modeled. All animats but animat ζ are built without feedback motors, which means that the current motor state can not be causal for future states, e.g. if at t_i the motors are on, the animat with feedback motors will know about it at $t_i + 1$.

To talk about the sensor-motor-model of the animats: The sensor for the wall returns 1 if the cell in front of the animat is a wall, otherwise, the sensor returns 0, respectively for the sides left and right. The sensor for the agent returns 1 if at the cell in front of the animat another animat is positioned, otherwise, the sensor returns 0, respectively for the sides left and right. The movement model contains four different cases mapped by a two bits tuple:

- $(0, 0)$ – no movement,
- $(0, 1)$ – turn right on the place,
- $(1, 0)$ – turn left on the place and
- $(1, 1)$ – to move one unit forward.

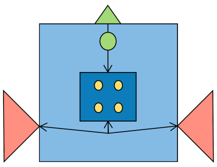
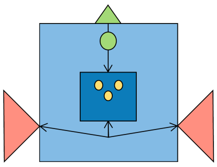
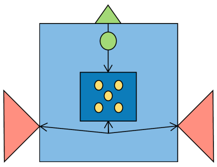
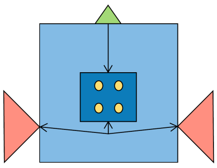
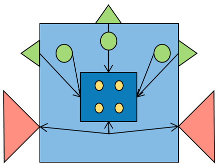
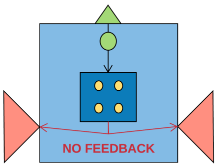
Name	α	β	γ
Architecture			
Hidden Nodes	4	3	5
Sense Animats	Yes	Yes	Yes
Sense Sides	Front	Front	Front
Name	δ	ϵ	ζ
Architecture			
Hidden Nodes	4	4	4
Sense Animats	No	Yes	Yes
Sense Sides	Front	Left, Front, Right	Front

Table 3: Six different animats were evolved in above world, they differ in the number and types of input sensors as well in the number of hidden sensors.

For the calculation of *cei* it matters, which nodes in the system can have causal power and similarly effective power. In this setting sensors can only be causal but the future states of sensors will affect the animats' performance. Hidden units should develop *cei*. As a special role in this settings animats $\alpha - \epsilon$ have motors with causal powers, too. This means that the current motor state could affect the future motor state. Only animat ζ has motors without causal powers.

4.3 Fitness Function To Evolve Collective Behavior

The problem in the simulated environment is multi-objective. On the one side, the animat receives a reward if it is able to travel through the gate (ref. figure 7) on the other side the animat gets a penalty if it hits another animat. Since it is much more likely that there is a collision between two animats, the penalty is lower as the bonus for traveling through the gate. If the penalty would be too low, all animats would center around the goal area and hit each other all the time. If the penalty would be too high, all animats would stop moving. Additionally, since it is not desirable that animats are moving condensed around the goal area, to receive reward for every single step they make, there is a timeout of 100 time steps until an animat can receive further points. This promotes also an evolution of a unified behavior.

In the following, a mathematical definition of the fitness function is given. In table 4 all necessary mathematical notations are explained. In equation (1) the fitness for a single animat in the environment is defined. In equation (2) the average fitness for all animats in the system is calculated.

a	Identifier of a single animat a, where $a \in \mathbb{N}$
A	The set of all animats a in an episode
$f(a)$	Returns the fitness of a single animat a
$F(A)$	Returns the fitness of the overall group A
$g(a, t_a, t_b)$	Returns the count of goals between time t_a and time t_b for a single animat a
$c(x, y)$	Returns the count of animats at a specific position (x, y)
t	A single time step t, where $t \in T$ and $t \in \mathbb{N}$
T	Set of all time steps t
$x(a), y(a)$	Returns the x and y position of animat a

Table 4: Definition of the Mathematical Notation for the Fitness Function

$$f(a) = \sum_{t=0}^{|T|/100-100} \begin{cases} 1 & g(a,t,t+100) > 1 \\ 0 & otherwise \end{cases} - \sum_{t=0}^{|T|} \begin{cases} 0.075 & c(x(a),y(a)) > 1 \\ 0 & otherwise \end{cases} \tag{1}$$

$$f(A) = \frac{\sum_{a=0}^{A} f(a)}{|A|} \tag{2}$$

4.4 The Communication and Cooperation of Animats

In this environment, there is no active collective behavior but rather a passive one. Due to the complexity of generating animats, which learn to communicate with each other, a decision was made to give animats only the knowledge about their neighbors (or not). This means that the agents do not share information actively, like two organisms making a dialog. Here the animats only receive the knowledge if there is another animat in its vicinity. This means that a group of animats does not have a collective behavior by definition. Nevertheless there should be evolved collective behavior, observable by a third person.

4.5 The Workflow of the Evaluation Procedure

Finally, it is explained, how the evolution procedure is modeled. It is referred to figure 8 on page 37 to describe the evolution process, additionally, the following paragraphs will provide detailed comments:

After initializing the world, clones of the original animat are generated for the current genome. In one evaluation run, only a single genome is evaluated, not the whole population. This takes much longer but since the performance of an animat is dependent on the position of the animat in the world. A really bad animat could win against a really good animat, if it is located near the goal area and the usually good animat is located somewhere in the corner.

The lifetime of the animats is 500 time steps (the duration of one episode). In each time step for every clone, the input values are determined and set. After updating the brain, which interprets the current states and calculates the future state, the motor states are received. There are four different actions: *Turn left, turn right, move forward* and *no movement*. According to the output, the position of the animat is updated as well as its performance. Dependent on the experiment setup the outputs are set to zero to prevent causal powers for output nodes. Finally, the average group score is calculated and returned to the optimizer. The fitness function above is used by the optimizer to assess the performance of a unique genome.

4.6 Required Data Analysis

There are many options to select routines and metrics for the evaluation of the integrated information. This section describes, which values were selected for the current model and how they are calculated.

Since calculations of Φ are computationally complex, only the animat with the highest fitness for each 100 generations is evaluated. In a world, with 10,000 generations there are 101 genomes (also the generation at time step zero is respected) to evaluate per test run. To flatten the effect of the random values in the GA, each experimental setting is repeated 30 times.

During the lifetime of an animat, it will not enter all possible states according to the *Transition Probability Matrix (TPM)* for the states. For example, an animat with sensors receiving only zeros would never experience all possible states. That is why all visited states are tracked during the lifetime of all clones. Afterwards, only for these states values of Φ are calculated.

In figure 9 on page 38, the abstract workflow for the evaluation process is given: For all finished experiments the genomes are extracted. For the elite subset of genomes, the simulation is executed again to track visited states and receive the TPM and the *Connectivity Matrix (CM)*. In this evaluation there are five different measures used, according to IIT [Alb+14]:

- $\langle \Phi^{Max} \rangle$: The quantity of integrated information in the main complex (MC). The MC is the irreducible set of mechanisms with the highest Φ.
- $\langle \Phi^{Max}_{Concepts} \rangle$: The number of concepts in the MC.
- $\langle \Phi^{Max}_{Elements} \rangle$: The number of mechanisms in the MC.
- $\langle \sum \varphi^{Max} \rangle$: The sum of the maximum φ of each mechanism.
- $\langle \sum \varphi^{Max}_{Concepts} \rangle$: The average number of concepts in the sum of the maximum φ of each mechanism.

For the selected measures the average values over all states, the animat is observed in, should be calculated as well the maximum value. Average values are marked by the symbols $\langle$ and $\rangle$.

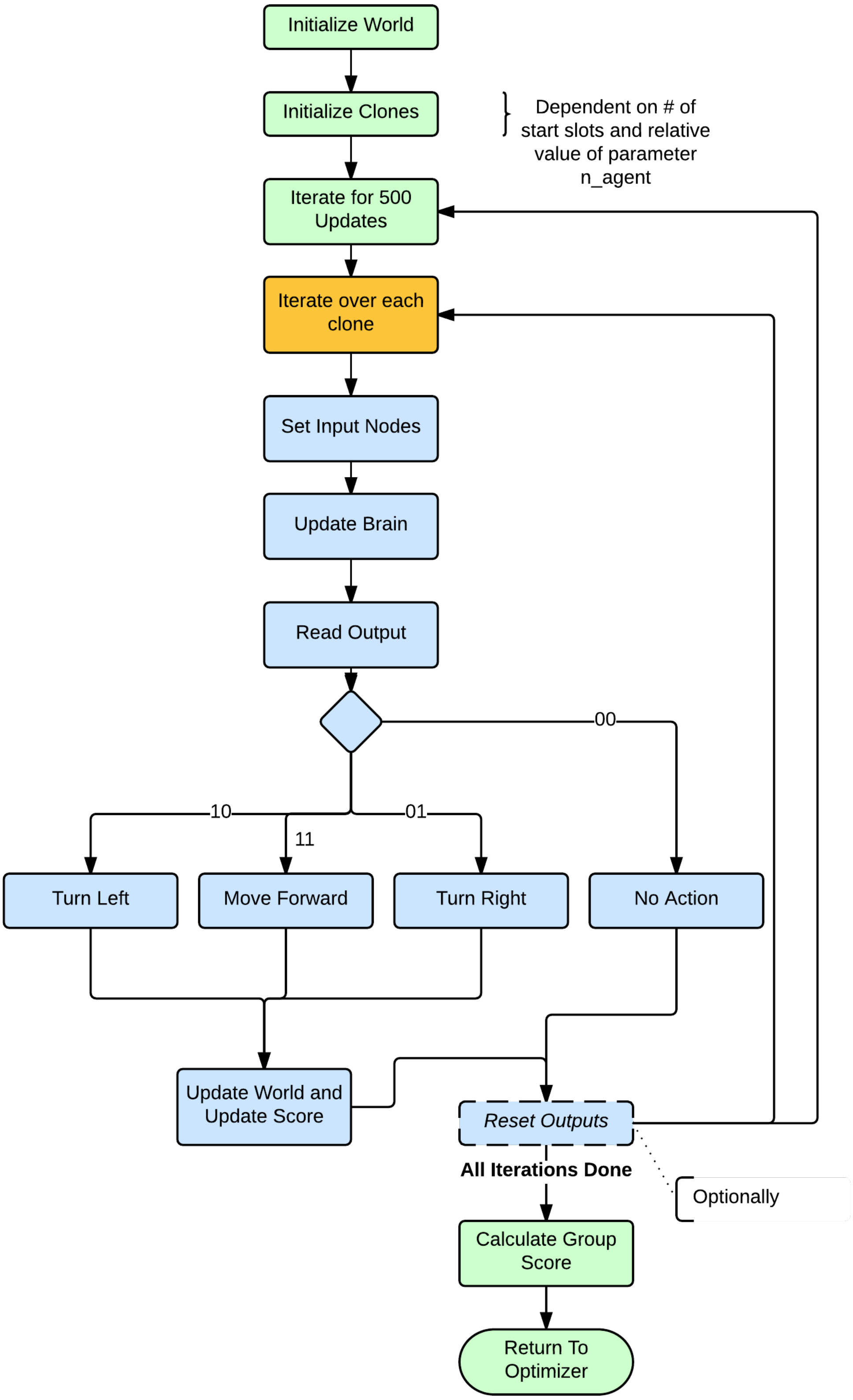


Figure 8: Workflow of one generation in the evolutionary process. The colors show the different depths of the iterations. Blue is the level of a single organism, yellow the level of one time step and green the level of one episode.

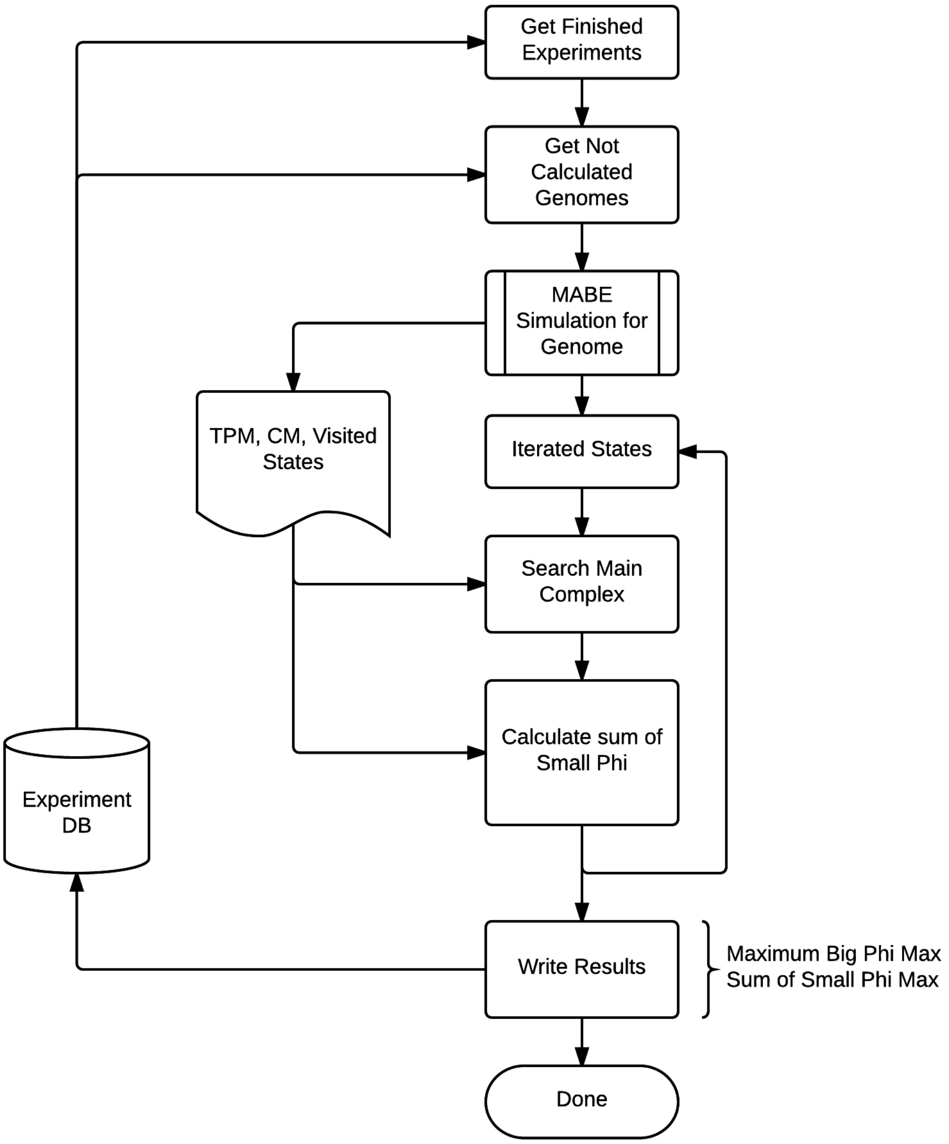


Figure 9: Abstract Process of the Evaluation Model

5 Implementation of the Evolution Model

Using the model for evolving collective behavior, an implementation was developed. An existing framework for multi-agent based evolutionary optimization, which supports Markov brains, was customized for this purpose. Furthermore multi-threading routines were developed, which helped to speed up the computation for the execution of the experiments and the implementation.

5.1 Evolving Animats with Integrated Information

Theoretically, it is possible to map a recurrent network, which is integrated, to a feed forward network. This would mean, that organisms with integrated information would have the identical behavior as organisms without. But if there is a limited number of gates and nodes available, integrated systems are required to get reasonable performance [OAT14]. That is why it is important to limit the capacities of the systems to force the EA to evolve integrated information in the animat's brain. Since the search space in Markov brains is multi-dimensional and an optimal search would require a lot of time, heuristics are used to evolve organisms with sufficient performance.

It is common to use evolutionary algorithms to develop collective behavior of artificial agents. In example works by Miikkulainen et al. used a framework called *NERO* [Mii+12]. *NERO* is a framework to develop intelligent agents. Furthermore, it has been shown that neuroevolution can be used to construct and promote even complex behaviors in homogenous and heterogenous teams [KJM15]. A further work, that supports the decision to use GA in combination with Markov brains is written by Perez et al. [Pér+07]. In their work GAs and the Baum-Welch Algorithm are compared in the case of learning *Hidden Markov Models (HMMs)* for *Human Activity Classification.* Also related here is the work by Xiao et al. [XZL07], where the goal was to optimize HMMs by a GA for *Web Information Extraction.*

The *Multi-Agent Based Evolver (MABE)*[6] is a young framework, which evolves digital organisms, that are acting in predefined environments. It was used in this work to implement the previously presented model. Besides its simple and transparent structure, the main reason for using MABE is that it supports the usage of Markov brains. It is convenient to generate a TPM out of a Markov brain, which makes it also simple to calculate its integrated information. It might also be possible to find integrated information in *ANNs (Artificial Neural Networks)* but this is not part of this work and would require a

[6] MABE Framework: `https://github.com/ahnt/MABE/`

remodeling of calculation routines in IIT 3.0. MABE consists of seven functional parts (ref. figure 10):

1. **World**: The environment an animat acts in.
2. **Organism**: The representation of the animat, having personal attributes a genome and a brain.
3. **Genome**: The evolved genome used to generate the brain.
4. **Brain**: The brain, which processes inputs and generates outputs.
5. **Optimizer**: It optimizes the genomes after each generation and produces the next population.
6. **Archivist**: Responsible of tracking all necessary measures and values to reproduce the results.
7. **Group**: Bundles populations and allows different settings between different groups (not relevant in this work).

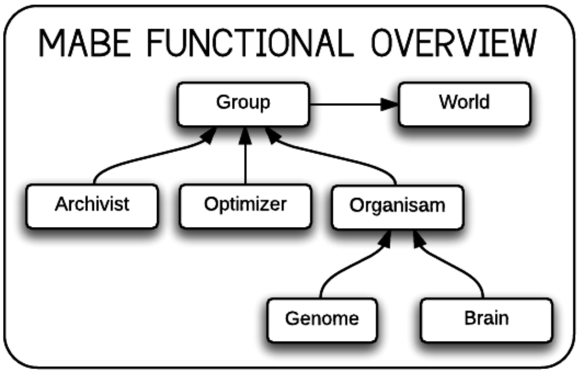


Figure 10: MABE functional overview[7], the world can evolve different populations (groups) of genomes. For each group an archivist is responsible for the data logging, an optimizer for evolving and optimizing the genomes and the population of organisms, which have to be evaluated in the world. A organism consists of his genome and his brain, the brain is constructed by the genome.

The evolution process was wrapped into the experimental environment (ref. figure 11). Before the animats were evolved the database was initialized. A *sqlite3*[8] database was used to store all data. Among them are all data, which were needed to handle the experiments and which were generated by the GA

[7] Image source: `https://github.com/ahnt/MABE/wiki`

[8] `https://www.sqlite.org`

and evaluation process. The Entity-Relationship diagram is explained in figure 12. After the setup process, the planned experiments are executed and their integrated information is measured. A configuration file helps to link database, experiment scripts and MABE. Review section A.1 on page I for further details.

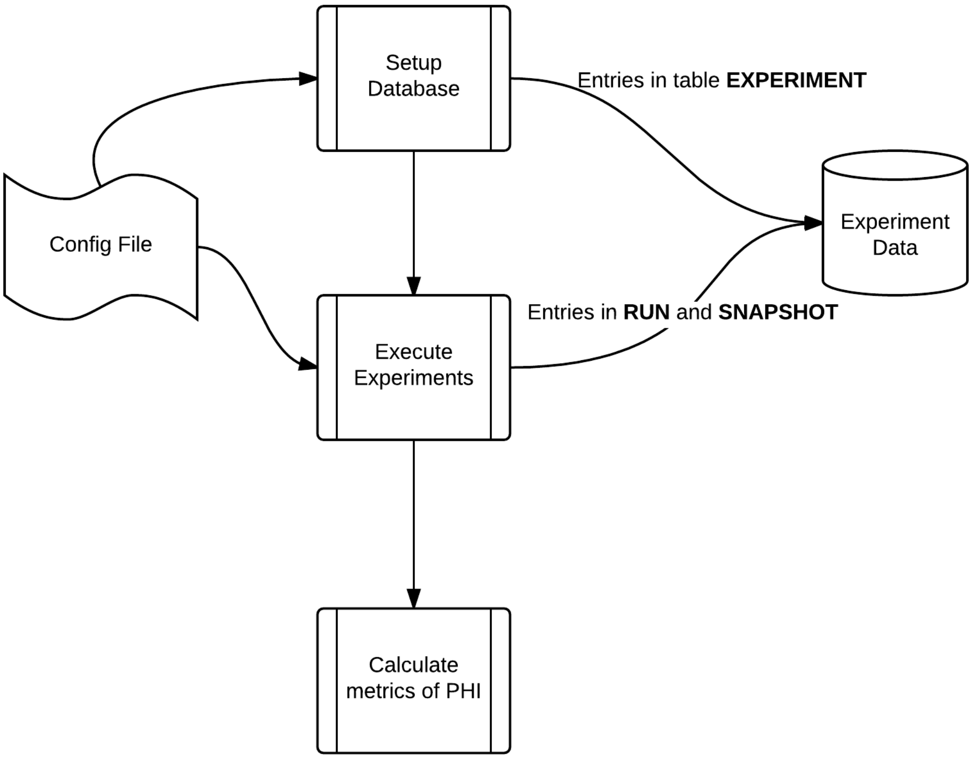


Figure 11: Basic Workflow of the Execution and Evaluation Process

For this work, the above-modeled environment, as well as the modeled animat, had to be implemented in MABE. Therefore a subclass of `World` and a subclass of `Organsim` was written. Besides all necessary routines for handling the existence of the animats and their behavior (direction, position) the fitness function was implemented in algorithm 1 on page 42, which updates the fitness after each step the animat makes:

Summing up all necessary experiments, there are 240 runs of the evolution process. In each run, there are 10.000 generations with a population of 100 genomes in each. For each genome, there is an episode over 500 time steps in the environment. This means that there are about 240,000,000 single test runs, which would last several weeks if running on a single core. That is the reason why a multi-threading approach was used to speed up that process. Follow figure 13 for the schematic architecture of this solution. Because there is no multi-threading implementation of MABE and on the other side there is a

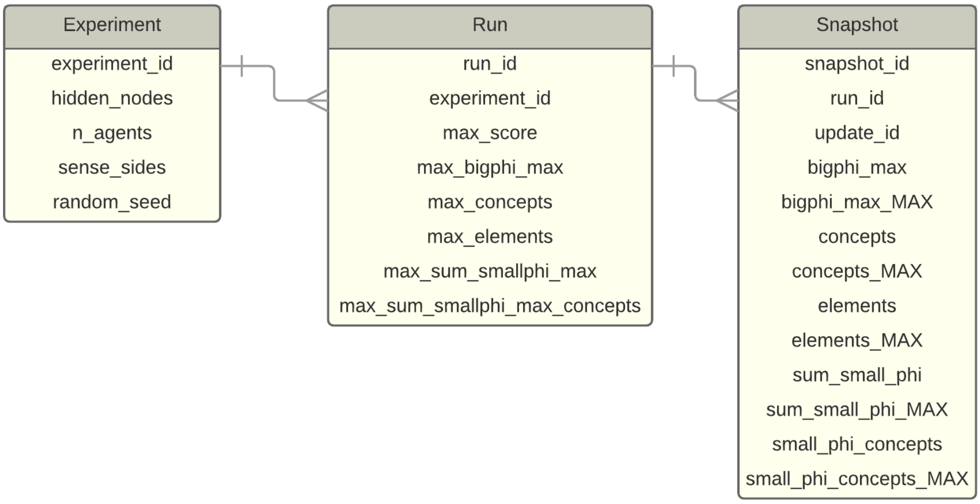


Figure 12**:** The experiment results were stored in a database. The *experiment* table describes the setup of a unique experiment. The *run* table describes a single executed experiment. The *snapshot* table describes a single snapshot in an experiment run. In table snapshot not only the average values of integrated information are stored but also the maximum values.

```
Data: newLocation, animat
Result: animat
$animat.waitForGoal = animat.waitForGoal - 1$;
if $newLocation <> animat.position$ then
    if $isGoal(newLocation)$ and $animat.waitForGoal <= 0$ then
        $animat.score = animat.score + 1$;
        $animat.waitForGoal = 100$;
    end
    if $isAgent(newLocation)$ then
        $animat.score = animat.score - 0.075$;
    end
    $animat.position = newLocation$;
end
```

Algorithm 1: How to Calculate the Score of an Animat

large count of experiments to run, whole experiments are warped into little packages to run them on different cores. This means that simultaneously there are multiple instances of MABE. Additionally, it is possible to split up the databases and distribute instances of the experimental environment to different machines to implement multi-computing. This makes the evolution process quite scalable. Unfortunately, there were issues replicating the animats using previously generated genomes and also issues calculating values of integrated information, which was due to the high count of nodes in some experiment settings. Depending on the experiment setting, between 85 and 99 percent of the genomes could be used for evaluation (ref. section A.2), but the high count of replications per experiment flattened this problem.

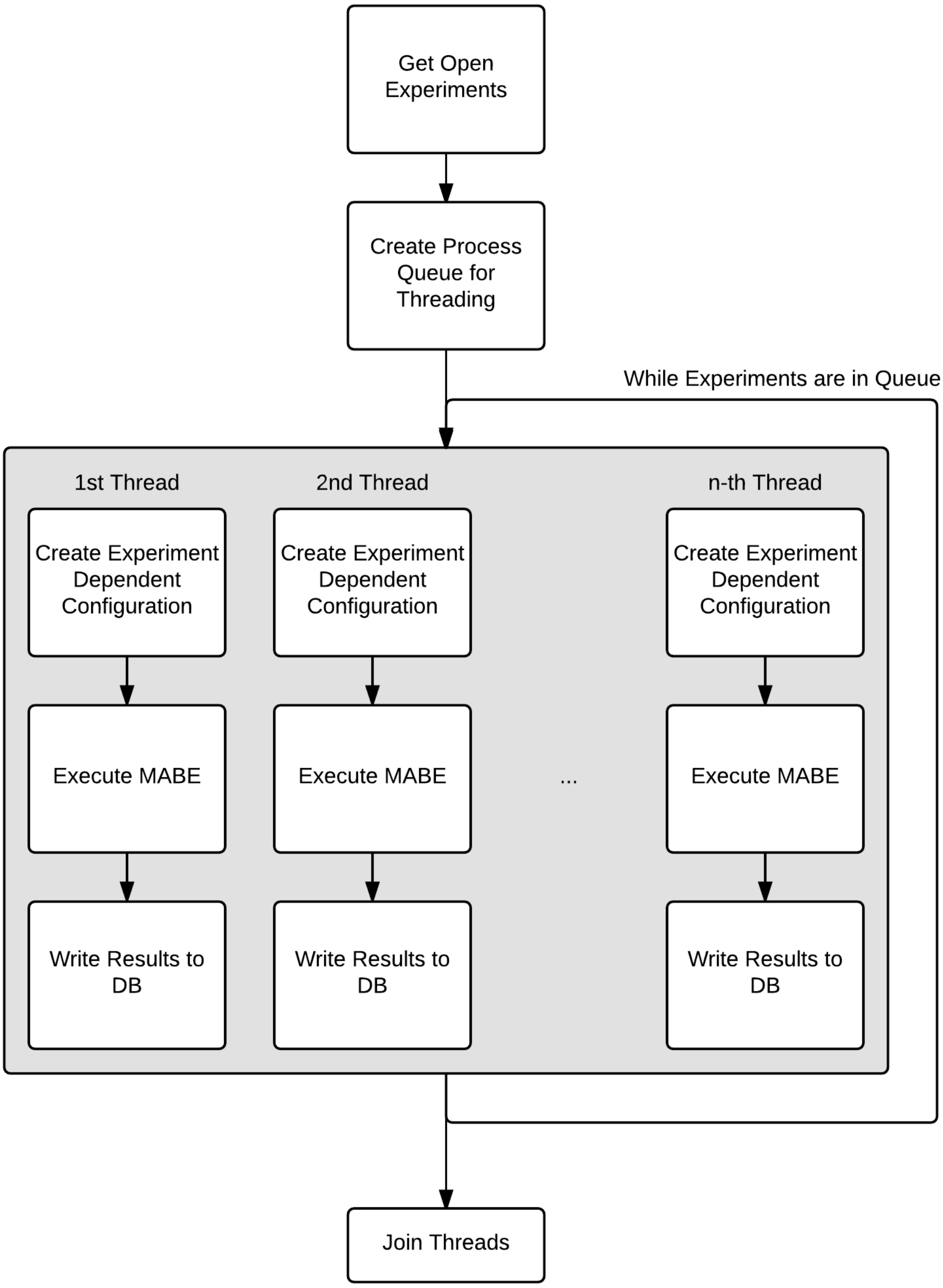

Get Open Experiments
Create Process Queue for Threading
While Experiments are in Queue
1st Thread
2nd Thread
n-th Thread
Create Experiment Dependent Configuration
Create Experiment Dependent Configuration
Create Experiment Dependent Configuration
Execute MABE
Execute MABE
...
Execute MABE
Write Results to DB
Write Results to DB
Write Results to DB
Join Threads

Figure 13: Multithreading is Used to Speedup the Overall Runtime

5.2 Calculating Integrated Information for Elite Organisms

Isolated from the optimization process, there is the calculation of the integrated information for elite genomes. An elite genome is the best genome of each population for a single generation. Since there are 10,000 generations in this settings and there are snapshots after each 100th time step, there are 101 genomes per run to evaluate. Depending on the number of causal mechanisms in a system the calculation of the values of Φ and φ can last several hours or even days. That is the main reason why only elite genomes are analyzed. For the calculations, a framework called *pyphi*[9], which helps to investigate integrated information. This framework was developed as part of IIT 3.0 [OAT14].

It is possible to calculate different values in IIT and make different aggregations of them. In algorithm 2 the calculation procedure is shown. Like in the model explained, five different measures are selected. Since the input mechanisms are never integrated into the system, the visited states have to be filtered before calculation. In the filtering process, all inputs are set to 0.5 (to avoid that they will be causal) and then duplicates are deleted. Due to performance issues, while calculating the values of integrated information, only the five most visited states are investigated. This is similar to the experiments in Albantakis et al. work [Alb+14]. For the calculation of φ power sets of the mechanisms need to be calculated to test the irreducibility of each possible set of mechanisms in a state. Finally, the values are written in the database.

[9] `https://github.com/wmayner/pyphi`

```
Data: CM, TPM, VisitedStates, nSensors, nHiddenUnits, nMotors
Result: $\langle\Phi^{Max}\rangle$,$\langle\sum\varphi^{Max}\rangle$,$\langle\Phi^{Max}_{Concepts}\rangle$,$\langle\Phi^{Max}_{Elements}\rangle$,$\langle\sum\varphi^{Max}_{Elements}\rangle$
/* For $\varphi$ and $\Phi$ only states of input and hidden nodes are
   considered                                                        */
$uniqueStates = filter(VisitedStates)$;
$currentNetwork = network(TPM, CM)$;
$valuesPHI = matrix()$;
for $state$ in $uniqueStates$ do
    $MC = findMC(currentNetwork, state)$;
    $valuesPHI.append([\Phi^{Max}, \Phi^{Max}_{Concepts}, \Phi^{Max}_{Elements}])$;
end
$ps_h = powerset(nHiddenUnits)$;
$ps_{hs} = powerset(nHiddenUnits, nSensors)$;
$ps_{hm} = powerset(nHiddenUnits, nMotors)$;
$valuesphi = matrix()$;
for $state$ in $uniqueStates$ do
    $subsystem = currentNetwork.subsystem(ps_h, ps_{hs}, ps_{hm})$;
    $concepts = subsystem.concepts()$;
    $sum_{phi} = 0$;
    for $concept$ in $concepts$ do
        $sum_{phi} + concept.phi^{Max}$;
    end
    $valuesphi.append([sum_{phi}, length(concepts)])$;
end
$results =$
$[valuesPHI.average(), valuesphi.average(), valuesPHI.max(), valuesphi.max()]$;

$database.write(results)$;
```

Algorithm 2: How to Calculate Measures of IIT

6 Results and Evaluation of the Experiments

The final contribution of this thesis is the presentation and interpretation of the simulated behavior and values of integrated information. First, it is described how the experiments were performed and then the actual results are listed. In GAs the resulting genomes and therefore also the resulting organisms are dependent on the random seed. This means, that for a valid evaluation it is necessary to repeat each experiment setting a reasonable amount of time to counter the random effects of the GA. To calculate reliable average values for each of the eight experiment settings the genetic algorithm run with thirty different random seeds. In each run, $10,000$ generations are generated with a population size of 100 genomes per generation. In the GA settings, a *Tournament optimizer* was chosen to optimize the organisms. This optimizer randomly selects two genomes and makes a mutation of the best performer to be added to the future population.

6.1 Experimental Setup

The *MABE* framework was configured with basic settings. Despite Markov brains could implement different types of HMGs[10] only deterministic HMGs were implemented. This means, that there are only two probabilities for a future state, zero and one. In future works also probabilistic types of gates should be considered.

All animats (α to ζ) were tested with *30 different random seeds.* Additionally, animat α was tested with *50 percent* and *75 percent* of the start slots available. According to the world design there are 72 start slots. If the experiment should run with *100* percent of the start slots 72 copies of the generated organism are created to be put in the environment. Resulting, eight different experiment settings are modeled.

Doing single evaluations of the genomes take a lot of time. Alternatively, it could be possible to implement group evaluation. This could enhance the performance a lot because it would not be necessary anymore to clone a single genome for all start slots. In group evaluation, all generated organisms are placed into the same world instance. Since the performance of an organism is depending on its position in the world, one animat could be rated better as a second one even if in general this would not be the case. A solution would be to repeat each generation a couple of times with different random positions for each animat to flatten this error. But since we wanted to eliminate as many

[10] Find a full list at `https://github.com/ahnt/MABE/`

random factors as possible, we decided to choose the isolated evaluation of a generated organism, by making clones of it and therefore only allow one type of genome in one episode.

6.2 The Evolution of Fitness Compared to Integrated Information

In the following pages, there are a bunch of plots showing the integrated information and fitness of all experiments (ref. figures 14 on page **??** and 15 on page 50). They should be reviewed during their explanations in the following paragraphs. The first plots show the architecture of the tested animats. This is followed by the *Lines of Decent (LODs)* for the relative fitness values. In all plots, the *Standard Error of the Mean (SEM)* is marked as an area behind the lines. The following five plots show the above-mentioned values of integrated information, also containing the SEM of the aggregated values.

In this model, it can be observed that Φ^{Max} is not as robust as $\sum \varphi^{Max}$. The integrated information of the whole system is present (systems are conscious), but the value has a high variance. On the other side, aggregations of the integrated information of single mechanisms φ have less variance and a higher correlation with the performance of the organism.

Comparing animat α and ζ, it can be seen that animats with feedback motors can achieve slightly better group performance and in average higher Φ^{Max}, which should be due to the higher count of possible causal nodes in the animats. But it is interesting to see that $\sum \varphi^{Max}$ stays more or less on the same level.

Having fewer animats in the collective causes a higher fitness. This is due to the fact, that if there are fewer animats in the environment the probability of hitting each other is lower. But the more interesting thing to look at is, that the mechanisms must be higher integrated to achieve such fitness. Because it is less likely to see a neighbor the animat needs more memory about the past (e.g. about the last occurrence of a neighbor) to perform well.

It turned out that varying the number of hidden nodes had no significant high impact in the animats' performance. While the performance (comparing α, β, γ) has the same growth the integrated information is more diverse. But on the other side, this can be good to make statements of how the number of hidden units in an animat influences the evolution of integrated information. It can be seen that Φ^{Max} has the best growth when using only 3 hidden units, while vice versa $\sum \varphi^{Max}$ is lower compared to animats with 5 hidden units, which is again due to the higher probability of integration of mechanisms in the larger system.

Using more or fewer sensors in the animats has clear effects on the performance, which was pretty much predictable before, but it is interesting how the integration was affected to this architectural modifications. To achieve at least some kind of a good performance, animat δ (with only one sensor) has to be highly integrated. On the other side for animats with lots of sensors, like animat ϵ who has six sensors, there is no need for a high integration and it is more or less a side-effect. The reason is that sensors can be mapped more directly to suitable outputs and there is no need for high memory in the system.

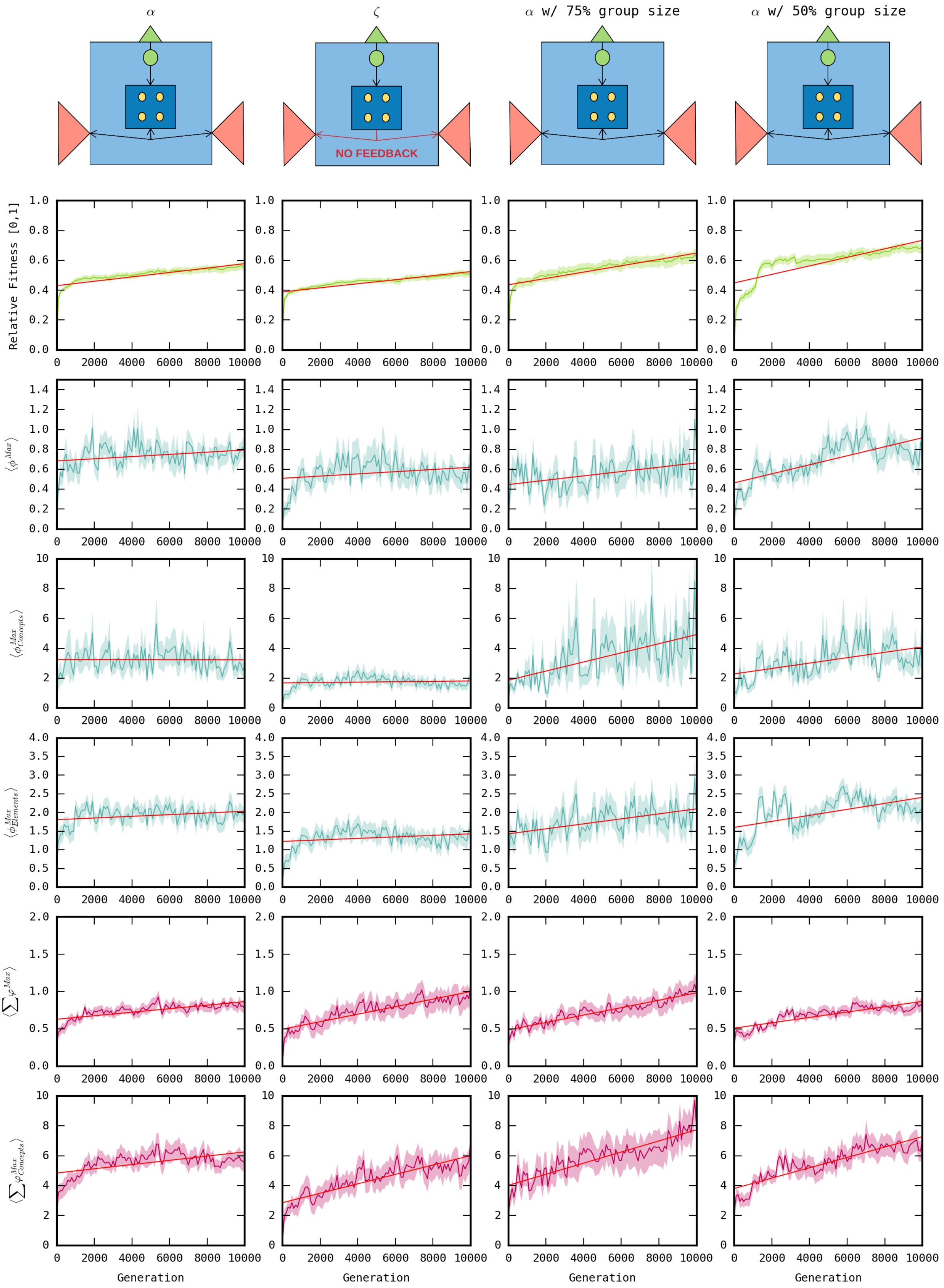


Figure 14: Experiments A. From top to bottom the figure shows the architecture of the animat, the relative fitness and the different IIT metrics. The area behind the lines show the SEM.

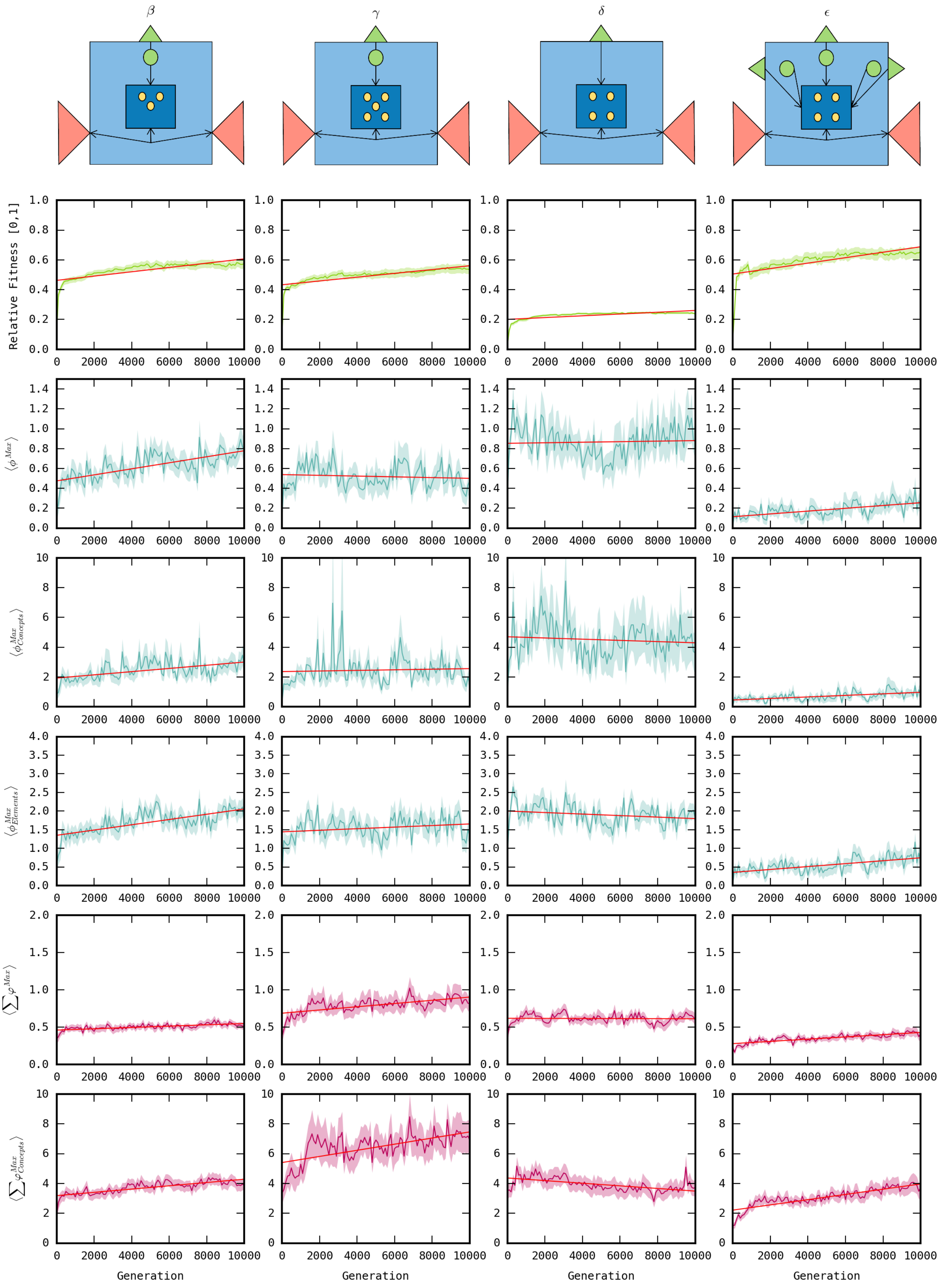


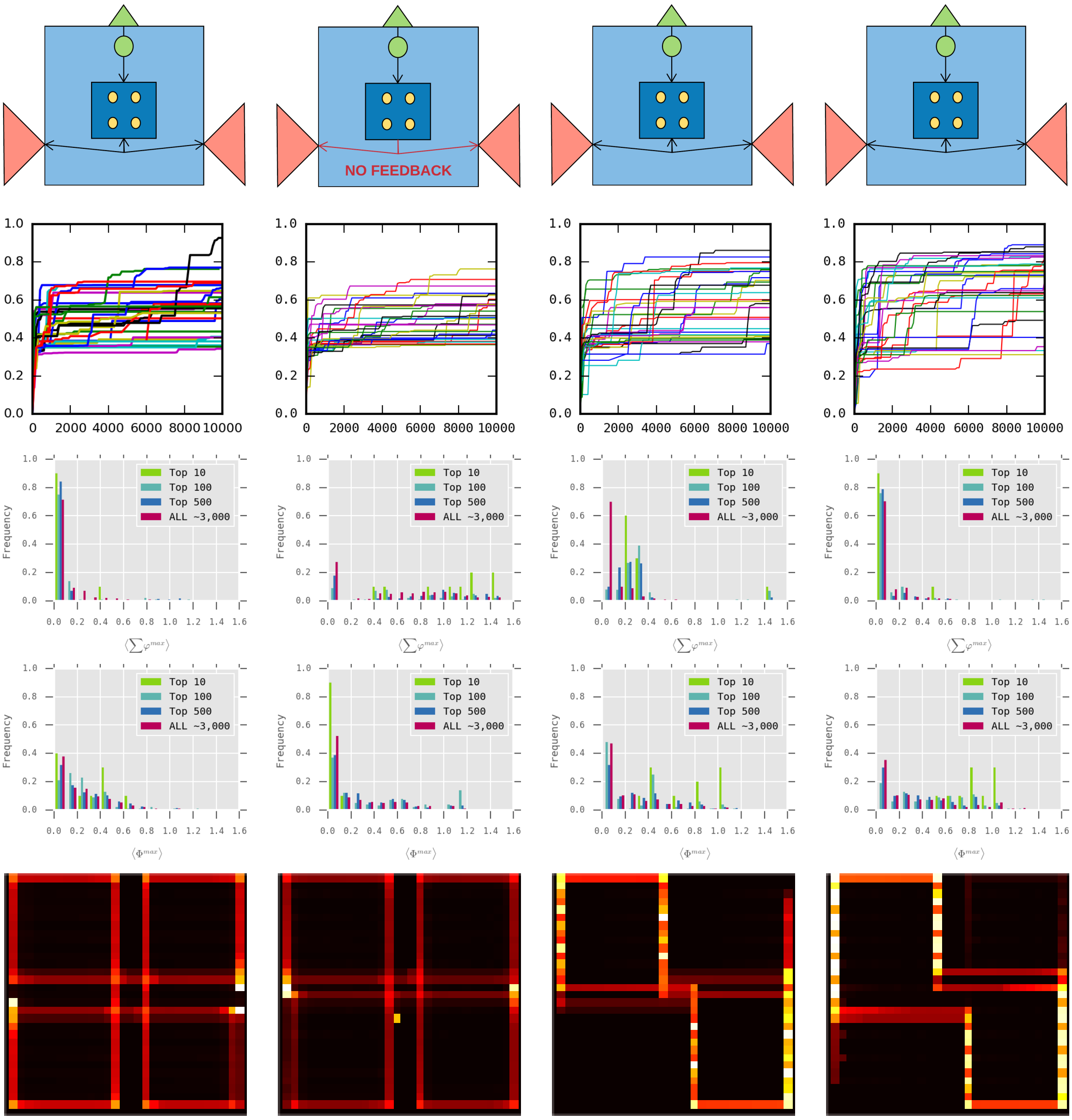
Figure 15: Experiments B. From top to bottom the figure shows the architecture of the animat, the relative fitness and the different IIT metrics. The area behind the lines show the SEM.

6.3 Further Insights on Behavior and Integration

In the tables 5 on page 53 and 6 on page 54, there are further plots for evaluation listed. First, there are all LODs for the fitness overall tested random seeds (bigger plots are in the appendix in section A.3). It can be seen that there is a high variance, but the *Standard Errors of the Mean (SEM)*, which are marked in the plots in figures 14 and 15, are reliable. Furthermore, there follow two histograms per animat comparing the $\langle\sum\varphi^{Max}\rangle$ and $\langle\Phi^{max}\rangle$ in four different categories: The *top-10* performers, the *top-100* performers, the *top-500* performers and all measured animats. This is followed by heat maps, which simply show the movement patterns of the collective: Black means that no animats visited this cell, white means that many animats visited that cell and red means something in between. An overview of heat maps over all best animats of an evolution can be found in section A.4. Finally, the values of information integration for the best performing animat per experiment are listed. In this section, there are statements about the content of these diverse plots.

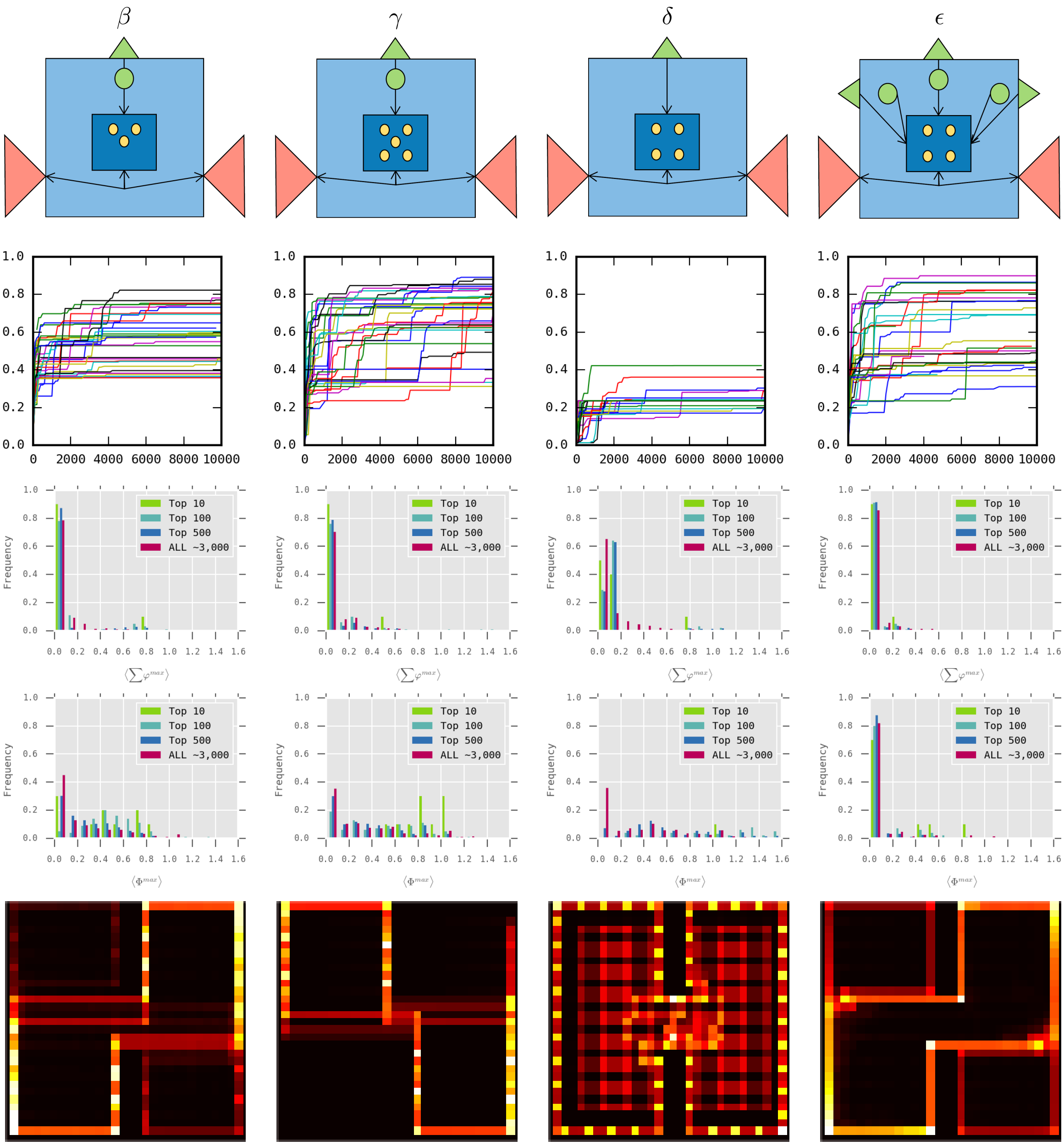
Considering only the best genomes of all different runs, in average it is more likely that genomes with a high performance also have a high integration, which can be observed in the histograms. But it is strongly related to the animat's architecture how this relation looks like. It is clear that for animats with a low value of $\langle\sum\varphi^{Max}\rangle$ and $\langle\Phi^{max}\rangle$ (e.g. animat ϵ with 6 sensors) this effect is minimal while for other animats this effect is easily observable (e.g. animat β with 3 hidden units). Again it is interesting to observe the different group sizes as well as animats with feedback motors in that manner: Forcing the animat to have more memory, like for animat ζ, the integrated information Φ^{Max} is higher for the top performers On the other side top performers on animats with feedback motors have significantly higher value of Φ^{Max} as the average. Looking at the different group size, top performers develop a higher $\langle\sum\varphi^{Max}\rangle$. It can be assumed that such animats are forced to have higher memories as animats in bigger groups since they can use neighbors as a guidance less likely.

The heat maps show that in general animats are using the walls for guidance (ref. figure 7 on page 32 to review the level). But additionally they are using shortcuts, after hitting the gate in the center of the environment, good performing animats cross the level to the other side of the room (left or right) until hitting a wall. Comparing the group size, it can be seen that there is a more fluid movement if there are more animats in the environment (the transition between red parts and white parts is not choppy). In example, if the heat map of the 50 percent group is observed, the animats standing around much longer, probably searching for neighbors or guiding walls, while animats with 100 percent of the populations have only bottlenecks on the corners. Additionally, if considering animats ϵ with six sensors, it can be seen that they evolve the ability to avoid other animats by searching a way around (this behavior can be observed in the left and right corner areas). This density-based behavior is similar to behavior like army ants, which also have a density-based movement model [Rei+15].



α	ζ	α w/ 75% swarm size	α w/ 50% swarm size
$\max_{a\in A}(f(A))_a = 3.71$	$\max_{a\in A}(f(A))_a = 3.06$	$\max_{a\in A}(f(A))_a = 3.45$	$\max_{a\in A}(f(A))_a = 3.56$
$\max_{a\in A}(\sum\varphi^{max})_a = 1.54$	$\max_{a\in A}(\sum\varphi^{max})_a = 0.55$	$\max_{a\in A}(\sum\varphi^{max})_a = 1.53$	$\max_{a\in A}(\sum\varphi^{max})_a = 0.77$
$\langle\sum\varphi^{Max}\rangle = 0.70$	$\langle\sum\varphi^{Max}\rangle = 0.10$	$\langle\sum\varphi^{Max}\rangle = 0.58$	$\langle\sum\varphi^{Max}\rangle = 0.35$
$\max_{a\in A}(\Phi^{max})_a = 1.50$	$\max_{a\in A}(\Phi^{max})_a = 0.13$	$\max_{a\in A}(\Phi^{max})_a = 1.0$	$\max_{a\in A}(\Phi^{max})_a = 1.5$
$\langle\Phi^{max}\rangle = 0.34$	$\langle\Phi^{max}\rangle = 0.02$	$\langle\Phi^{max}\rangle = 0.33$	$\langle\Phi^{max}\rangle = 0.74$

Table 5: Evaluation of the Top Performers A. From top to bottom there are plots for the architecture of the animat, all LODs, histograms showing the integration in relation to the animats' performance, a heat map marking the behavior pattern of the best performer and the top values for information integration.



$\max_{a\in A}(f(A))_a = 3.29$	$\max_{a\in A}(f(A))_a = 3.57$	$\max_{a\in A}(f(A))_a = 1.69$	$\max_{a\in A}(f(A))_a = 3.60$
$\max_{a\in A}(\sum\varphi^{max})_a = 0.52$	$\max_{a\in A}(\sum\varphi^{max})_a = 2.34$	$\max_{a\in A}(\sum\varphi^{max})_a = 1.15$	$\max_{a\in A}(\sum\varphi^{max})_a = 0.19$
$\langle\sum\varphi^{Max}\rangle = 0.20$	$\langle\sum\varphi^{Max}\rangle = 1.03$	$\langle\sum\varphi^{Max}\rangle = 0.51$	$\langle\sum\varphi^{Max}\rangle = 0.08$
$\max_{a\in A}(\Phi^{max})_a = 2.11$	$\max_{a\in A}(\Phi^{max})_a = 1.0$	$\max_{a\in A}(\Phi^{max})_a = 5.94$	$\max_{a\in A}(\Phi^{max})_a = 0.22$
$\langle\Phi^{max}\rangle = 0.80$	$\langle\Phi^{max}\rangle = 0.36$	$\langle\Phi^{max}\rangle = 4.49$	$\langle\Phi^{max}\rangle = 0.04$

Table 6: Evaluation of the Top Performers B. From top to bottom there are plots for the architectue of the animat, all LODs, histograms showing the integration in relation to the animats' performance, a heat map marking the behavior pattern of the best performer and the top values for information integration.

6.4 On the Lack of Correlation Between Fitness and Measures of Integrated Information

To talk about the correlations (ref. table 7), between the fitness and the integrated information, it can be seen that there is no significant correlation between the fitness values and integrated information. But this has also its valid relations to IIT:

According to IIT, there might be physical systems with no behavior at all (no measurable performance), but with a high conscious experience [TK15]. Take patients with the lock-in-syndrome, who are totally paralyzed and not able to communicate with others, they would be conscious without observable behavior.

		$\langle\sum\varphi^{Max}\rangle_{Max}$	$\langle\sum\varphi^{Max}_{Concepts}\rangle_{Max}$	$\langle\Phi^{Max}\rangle_{Max}$	$\langle\Phi^{Max}_{Elements}\rangle_{Max}$	$\langle\Phi^{Max}_{Concepts}\rangle_{Max}$
α	$\langle R\rangle$	0.26	0.23	0.12	0.09	0.08
	SEM	0.07	0.08	0.06	0.06	0.07
β	$\langle R\rangle$	0.17	0.27	0.22	0.24	0.20
	SEM	0.08	0.08	0.06	0.06	0.06
γ	$\langle R\rangle$	0.24	0.35	0.07	0.09	0.10
	SEM	0.09	0.08	0.07	0.06	0.06
δ	$\langle R\rangle$	0.07	-0.01	0.00	-0.02	-0.04
	SEM	0.06	0.06	0.06	0.05	0.06
ϵ	$\langle R\rangle$	0.30	0.38	0.09	0.08	0.07
	SEM	0.06	0.05	0.07	0.05	0.05
$\alpha * 0.75$	$\langle R\rangle$	0.24	0.19	0.09	0.11	0.11
	SEM	0.08	0.08	0.05	0.06	0.06
$\alpha * 0.5$	$\langle R\rangle$	0.34	0.45	0.21	0.21	0.20
	SEM	0.07	0.06	0.07	0.06	0.06
ζ	$\langle R\rangle$	0.29	0.33	0.10	0.14	0.13
	SEM	0.07	0.07	0.06	0.05	0.05

Table 7: Average correlation coefficients $\langle R\rangle$ (Spearman rank) over all 30 unique experiments per setting. Compared are measures of IIT to the fitness.

On the other side, it is also valid that there are organisms with a high fitness and no Φ^{Max} at all. This can be explained by the argument by Searle, which was explained in section 3.1. We are only able to make implications about the consciousness by observing their behavior [Sea00]: *I know that I am conscious and therefore I imply that you are conscious.* Without measures of consciousness, we only would trust such implications if we would have to judge about a system is conscious or not. To bring that together, it is totally valid that there are top-performing animats with no integration and on the other side really bad performers with high integration.

6.5 Dissect of the Best Animat Brain

Exemplarily the best performing brain for animat α was picked for detailed investigations. The animat has a score of 3.71, which conforms a fitness of 92.75%. In figure 16, two graphs are shown. The left graph shows the wiring of the brain nodes to the HMGs. Blue nodes are sensors, yellow nodes are motors and green nodes are hidden units. Black nodes are HMGs. The top line of nodes view the Markov brain at t_i, afterward the HMGs translates the node values to the new state at t_{i+1}, which is shown on the bottom line. On the right side, there is the wiring diagram of the resulting system, which already looks rather integrated. In the graphs, it is easy to spot that in this case there can be no effects on sensors, but motors with causality. In figure 17 the same system is visualized in the state with the highest Φ^{Max} of 0.33 in the main complex. The current state causing this integrated information is $(0, 1, 0, 0, 1, 0, 0, 1)$. This means that the first sensor is zero and the second sensor is one, followed by two zero motors and the hidden units where two of them are active.

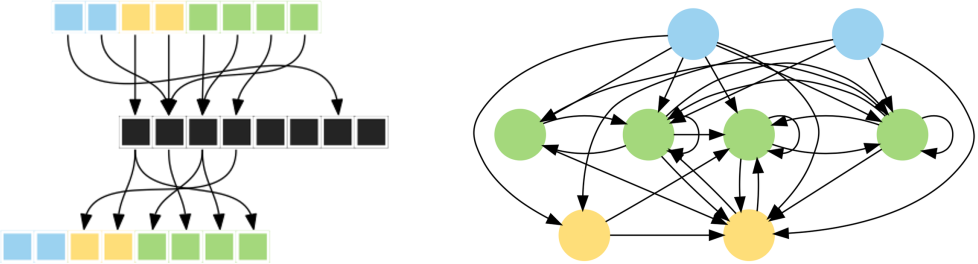


Figure 16: Wiring Diagrams of the Best Performing Genome of Animat α

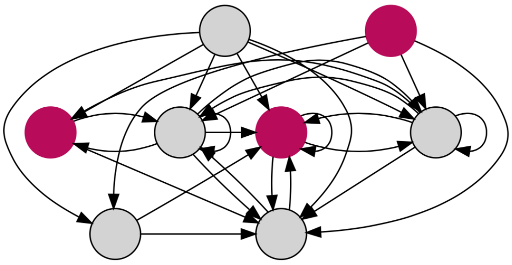
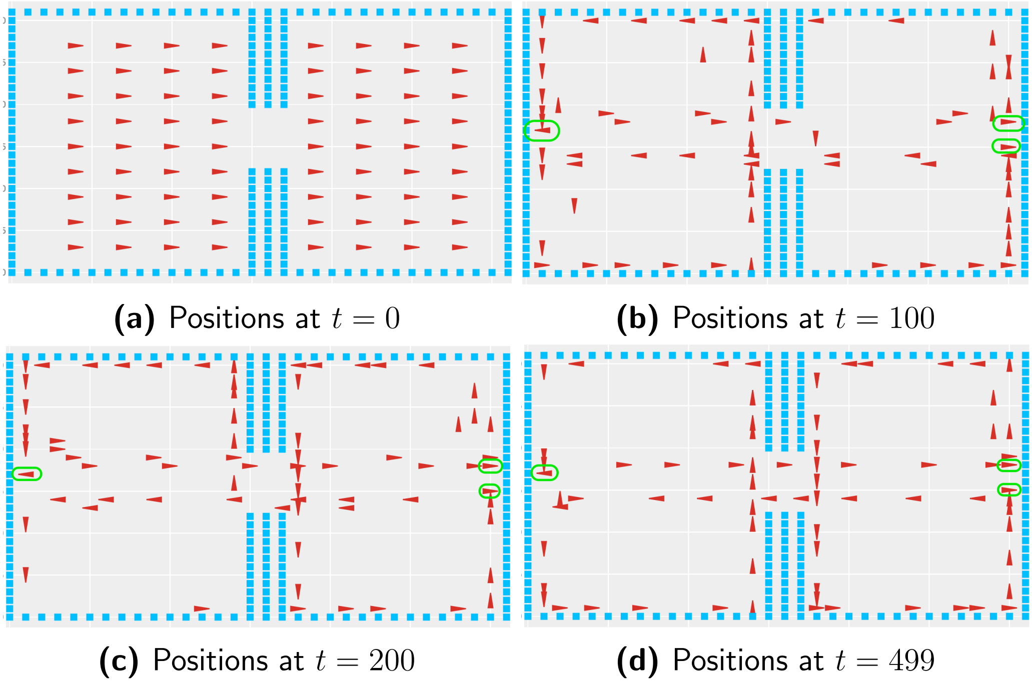


Figure 17: State With the Highest Φ^{Max} of the Best Genome of Animat α

Additionally, the positions of the group were visualized for further insights (ref. figure 18). The plots show the environment at four different time steps ($t = 0, 100, 200, 499$). At $t = 0$ all animats are in their starting position. In the following time steps, the animats are simply following the walls and trying to cross the gate in the middle. What is hard to spot in static plots but observable



(a) Positions at $t = 0$

(b) Positions at $t = 100$

(c) Positions at $t = 200$

(d) Positions at $t = 499$

Figure 18: Collective Behavior of Animats Visualized

in the animation[11] the behavior of the animats, which are marked with green circles (ref. figure 18 b - d): Once hitting the wall, they stay static and will never receive any points. But on the other side that static animats serve as guides for the other animats. The moving animat will turn left when it sees this the animat in the front of itself. In other *(naive)* words it can be said that the static animats sacrifice themselves for the better performance of the collective. It can be discussed if this effect is only by coincidence or by evolution. It would need further empirical research on the results to make robust statements about the cooperation of the animats and the evidence of similar selfless behavior.

[11] A movie clip can be found on the attached disk (ref. file `99_movie_A26.mov`)

7 Future Work and Discussion

To make further investigations on the behavior and to make it more difficult for the animats to have good performance, it is important to dynamically change the environment. Currently, the animats only need to follow the wall to achieve good performance. It would be harder if additional levels would be modeled where this rule would not hold. The reason for not allowing complex animats (with higher numbers of sensors and hidden units) was that there would only integrated information by coincidence because it would not be necessary to have a large memory about the past to make a good performance. Another reason is that the calculation of integrated information is currently done without heuristics. This causes excessive runtime in larger systems. Before investigating larger systems, alternative algorithms should be found to make efficient calculations. Another extension to the model would be the simulation of real world organisms (e.g. [CF03; Rei+15; INI10]). This could help comparing insights of simulated machine consciousness to their real life examples.

Currently, each genome is evaluated independently by cloning it many times to get a homogenous group of animats. This consumes a lot of performance in the optimization process of the GA. To solve this issue it should be worked on solutions where a group of heterogeneous genomes are evaluated simultaneously. It possible requires more generations until high performance is reached but the evolutionary process will run much faster. Additionally, the optimizer values should be tuned. At the moment the variety of the different replications of the test runs is pretty high. Depending on the random seed, either good or bad results are received. Working on the right optimizer should shrink this problem. Additionally, alternative HMGs could be tested. In this thesis, only deterministic gates are used but there are also alternatives like HMGs using neural networks or probabilistic HMGs.

There can be also many variations on the calculations of the integrated information. For this work, the procedure used by Albantakis et al. [Alb+14] was rebuilt. Also, like in the work by Albantakis et al., due to performance reasons, only the five most visited states were used to calculate the integrated information. For further works in this field, a unified model of calculating integrated information in animats would be the base for homogeneous comparisons.

Most of the resulting statements in this work were done by aggregating the different experiments. There should be more research done on the intrinsic structure of single animats to make further statements about their behavior. Currently, only one best performing animat was presented, but to find universal statements, such investigations need to be done for a reasonable number of animats. Furthermore it would be a special task to question if a whole collective can evolve integrated information if it is seen as a single system.

8 Conclusion

Usually, discussions about consciousness and machine consciousness can last over years or even decades without getting to an end. People might be right arguing that IIT is based on facts that are not proved yet. [Cer15] But even assuming that IIT is a faulty theory and be able to prove that would contribute exceptionally to the research of consciousness. The same statement would be valid for research on machine consciousness. Finally, proving that it is not possible to put a soul into a machine would be a breakthrough in this scope. A message, which came up working on this thesis is, that the high diversity of different theories and opinions should be respected and not be prejudiced.

Puzzling on the big problems *of strong AI, P-Consciousness, super-intelligence,* etcetera will gain importance in relevance in the next years. This work should be a beginning approach to see the problem not only isolated to single systems (e.g. machines, humans, dogs) but to work on the consciousness of collectives. This would mean to investigate if a whole swarm of animals would have integrated information, the internet or even systems like cities. It is undeniable that everything influences each other, that is why more work should be invested the research of consciousness of collectives.

In this work, it could be shown that animats with collective behavior are more likely to have integrated information than others. Additionally, it can be seen that integrated information is dependent on the size of the group working together and of course on the complexity of the brain's architecture. Despite there is no active communication between the animats, neighbors are used as a guidance. It was even a scenario investigated where single animats sacrificed themselves to have a better average performance.

This thesis can be closed with suitable solutions to the defined objectives: A broad overview of the research in consciousness and machine consciousness was given. Using accepted philosophical and cognitive theories of machine consciousness a model was developed to evolve artificial animats with simulated consciousness and collective behavior, which could be implemented successfully later. Finally, suitable and valid statements could be made about the integrated information of the evolved animats.

A Appendix

In this section a explanation over the used scripts is given as well as more detailed and additional plots.

A.1 Scripts Used for the Experiments

Python scripts to setup the database, run the experiments and visualize movements:

1. `best_HM.py`: Plots a heat map for each experiment setting, visualizing the movement of the best performer.
2. `calculate_phi.py`: Calculates the values of integrated information for the genomes of a single evolution.
3. `execute_simulation.py`: A batch script to execute a set of planed experiments.
4. `make_heatmaps.py`: Plots all heat maps of a experiment settings containing the best performers of each random seed.
5. `setup_db.py`: Initializes the sqlite database.
6. `start_bigphicalc.py`: Manages the calculation of integrated information.
7. `update_db.py`: Adds columns to the database.
8. `update_db2.py`: Changes the type of columns in the database.
9. `update_db3.py`: Adds an index to the table `snapshot`.

Jupyter Notebook[12] scripts for the evaluation and visualization of the experiments:

1. `01_FillPhiValuesToRun.ipynb`: Writes the values of the best performers per run into the table `run`.
2. `02_makePlots_LOD_A4.ipynb`: Plots the aggregated lines of decent.
3. `03_makePlots_all_LOD.ipynb`: Plots all lines of decent of the fitness per every experiment setting.

[12] `http://jupyter.org`

4. `04_makePlots_Corr_Fitness-vs-PHI.ipynb`: Calculates the correlation coefficients between the fitness and values of integrated information.

5. `05_makePlots_SCORE_VS_BIGPHI.ipynb`: Plots the histograms comparing top performers and ϕ^{Max}

6. `06_makePlots_SCORE_VS_PHI.ipynb`: Plots the histograms comparing top performers and $\sum \varphi^{Max}$

7. `07_visualize_network.ipynb`: Generates an animation for the best genome of a selected evolution.

A.2 Quality of Evaluations

Due to issues in MABE with replicating the brain using genomes, not all genomes could be evaluated. Additionally for the complex experiment settings β and ϵ the pyphi framework was not able to compute a subset of networks:

- Experiment setting α: 99.33%
- Experiment setting β: 94.75%
- Experiment setting γ: 84.30%
- Experiment setting δ: 96.86%
- Experiment setting ϵ: 84.45%
- Experiment setting $\alpha * 0.75$: 94.07%
- Experiment setting $\alpha * 0.50$: 92.33%
- Experiment setting ζ: 95.9%

A.3 All Line of Decents

In the following all lines of decent from section 6.3 are listed in a better resolution.

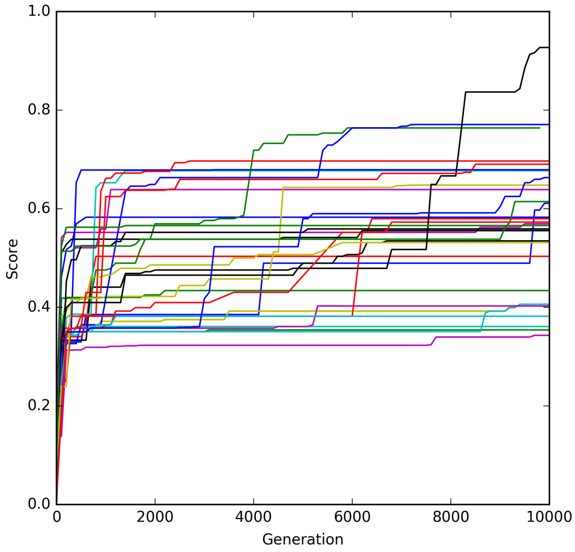


Figure 19: All LODs for Experiment Setting α

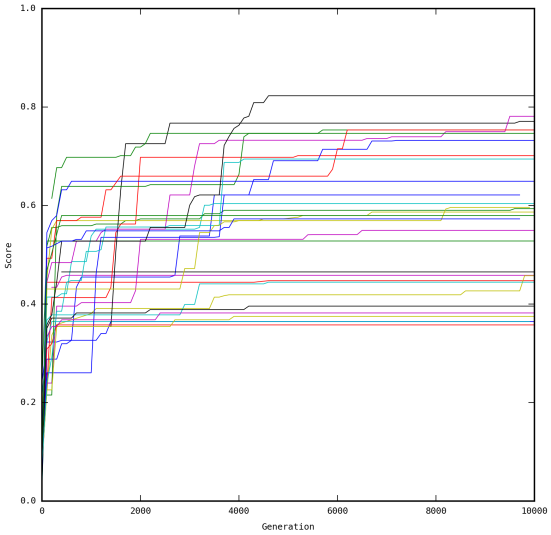


Figure 20: All LODs for Experiment Setting β

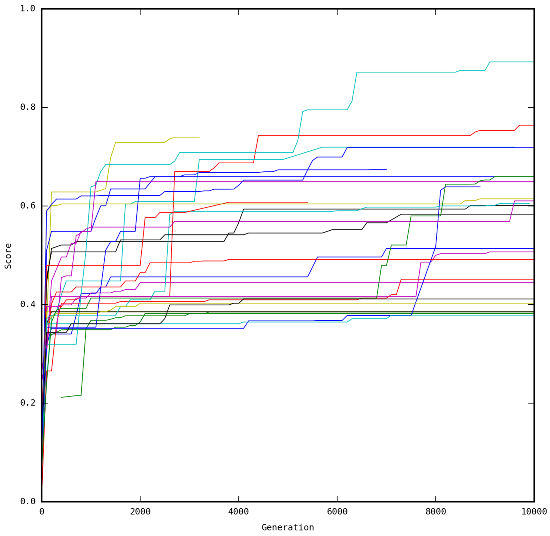


Figure 21: All LODs for Experiment Setting γ

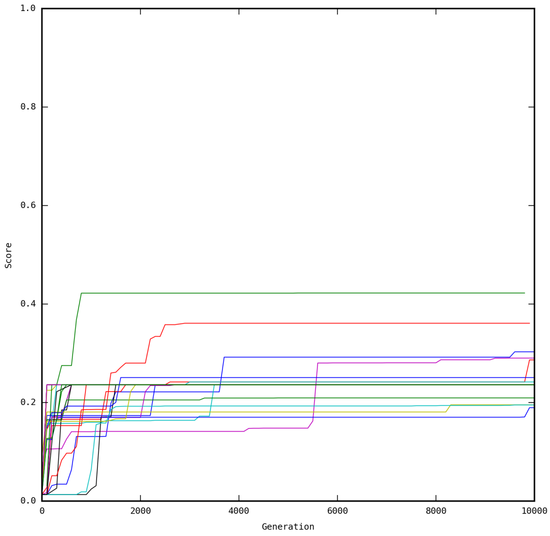


Figure 22: All LODs for Experiment Setting δ

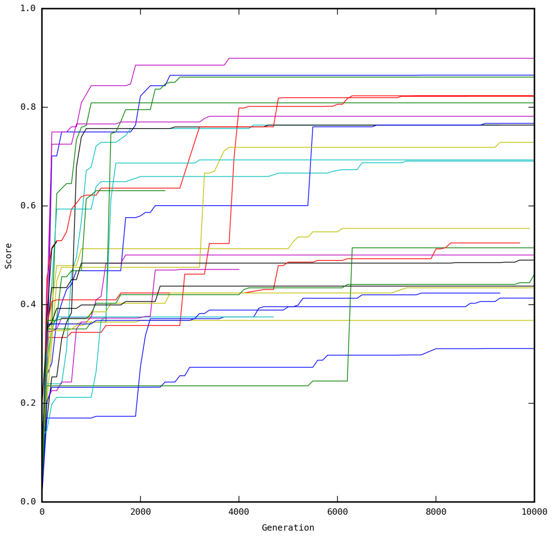


Figure 23: All LODs for Experiment Setting ϵ

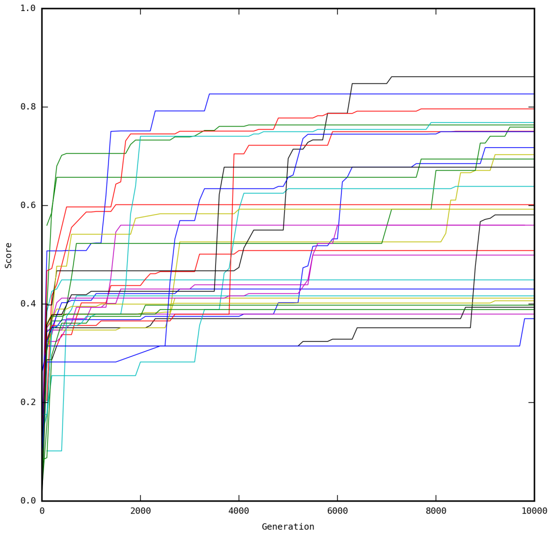


Figure 24: All LODs for Experiment Setting $\alpha * 0.75$

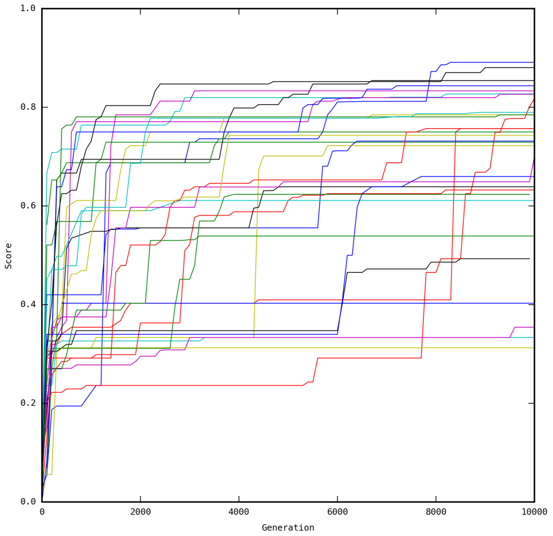


Figure 25: All LODs for Experiment Setting $\alpha * 0.50$

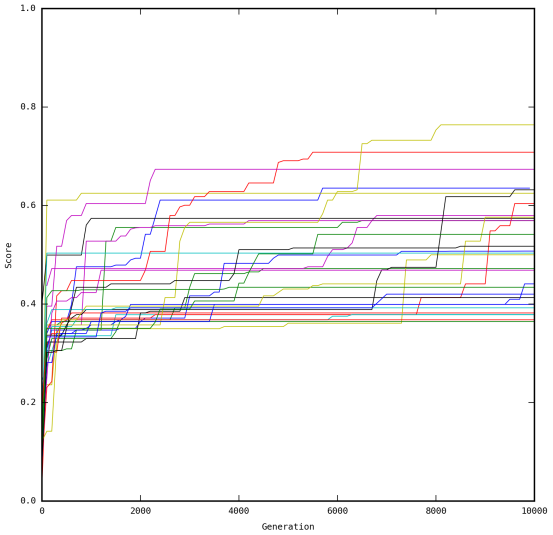


Figure 26: All LODs for Experiment Setting ζ

A.4 Further Heat Maps of all Best Performers

In the following heat maps of all best performers of every evolution are shown. The plots are grouped by experiment setting.

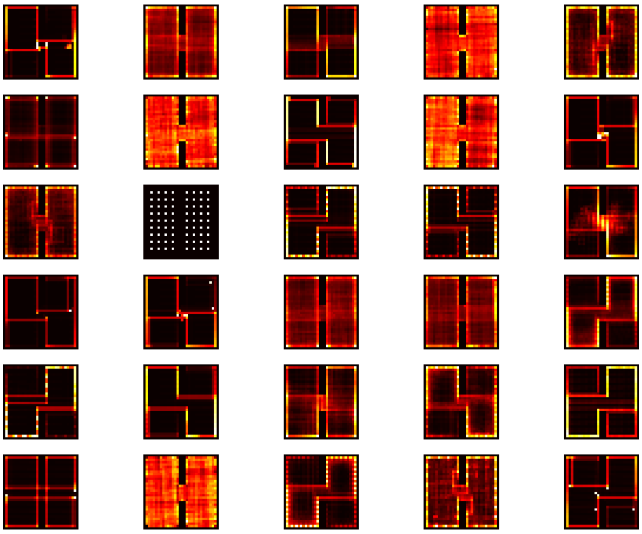


Figure 27: All Heat Maps for the Best Animat in an Evolution for Animat α

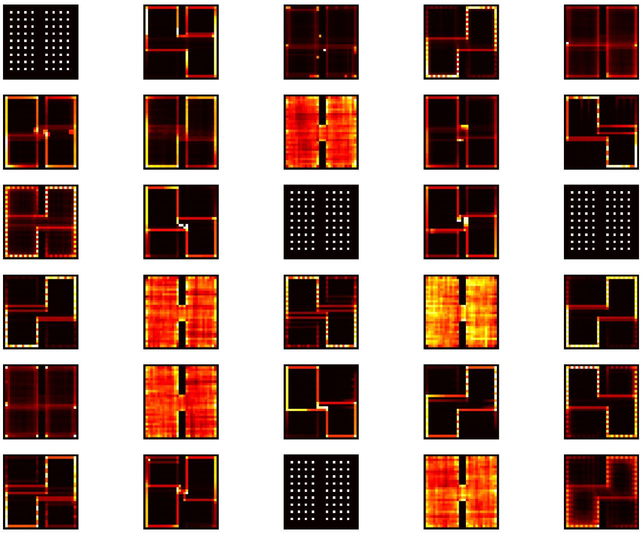


Figure 28: All Heat Maps for the Best Animat in an Evolution for Animat β

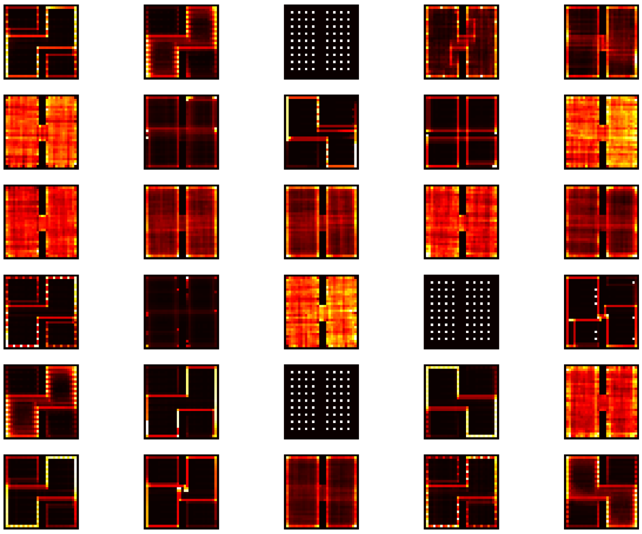


Figure 29: All Heat Maps for the Best Animat in an Evolution for Animat γ

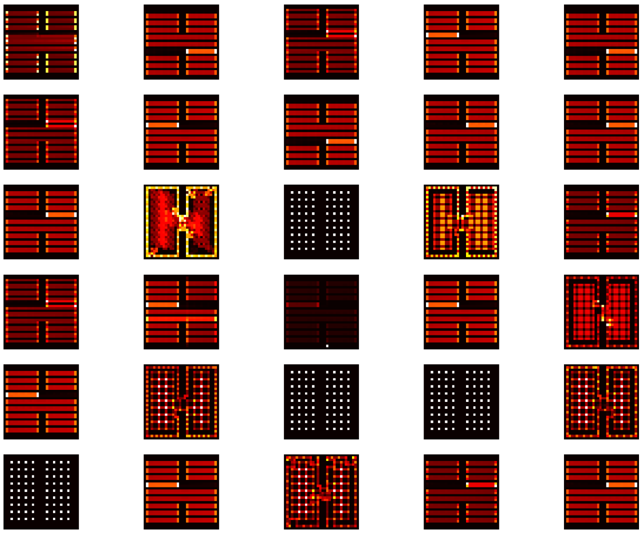


Figure 30: All Heat Maps for the Best Animat in an Evolution for Animat δ

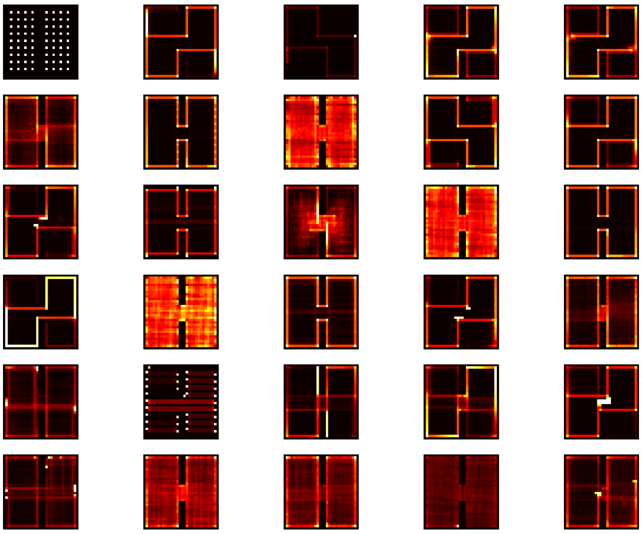


Figure 31: All Heat Maps for the Best Animat in an Evolution for Animat ϵ

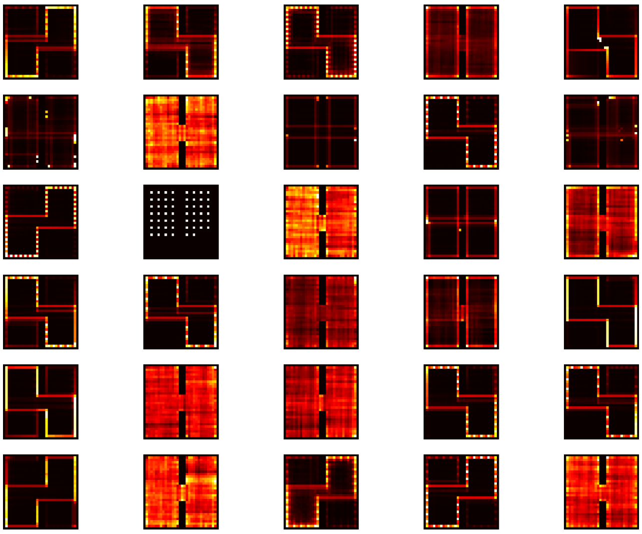


Figure 32: All Heat Maps for the Best Animat in an Evolution for Animat $\alpha * 0.75$

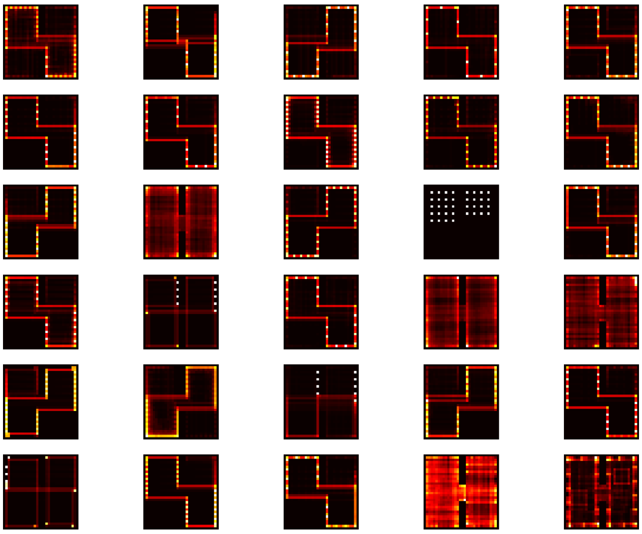


Figure 33: All Heat Maps for the Best Animat in an Evolution for Animat $\alpha * 0.50$

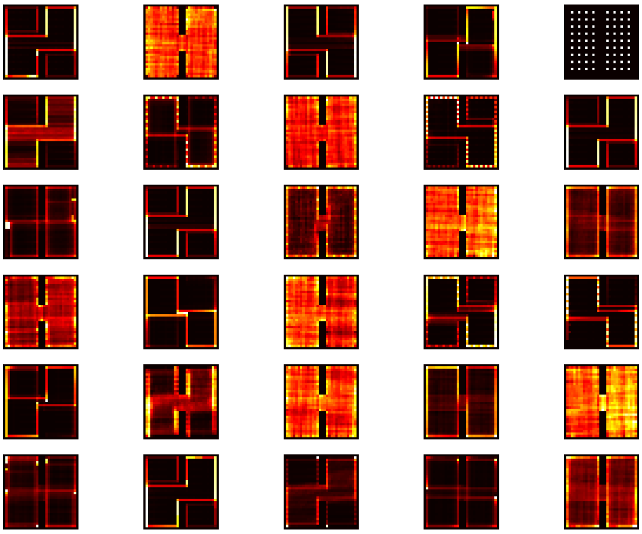


Figure 34: All Heat Maps for the Best Animat in an Evolution for Animat ζ

B Literature

[Alb+14] Larissa Albantakis et al. “Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity”. In: *PLoS Comput. Biol.* 10.12 (Dec. 2014). Ed. by Daniel Polani, e1003966. ISSN: 1553-7358. DOI: `10.1371/journal.pcbi.1003966`.

[Ann16] George J. Annas. “The Man on the Moon”. In: *Sci. Fict. Philos. From Time Travel to Superintelligence.* Wiley, 2016, pp. 245–259. ISBN: 9781118922620.

[Baa02] Bernard J. Baars. “The conscious access hypothesis: origins and recent evidence”. In: *Trends Cogn. Sci.* 6.1 (2002), pp. 47–52.

[Baa88] Bernard J. Baars. “A cognitive theory of consciousness”. In: *NY Cambridge Univ. Press* (1988).

[Baa97] Bernard J. Baars. *In the Theater of Consciousness.* 4. Oxford University Press, Mar. 1997, pp. 292–309. ISBN: 9780195102659. DOI: `10.1093/acprof:oso/9780195102659.001.1`.

[BF07] Bernard J. Baars and Stan Franklin. “An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA”. In: *Neural Networks* 20.9 (Nov. 2007), pp. 955–961. ISSN: 08936080. DOI: `10.1016/j.neunet.2007.09.013`.

[BF09] Bernard J. Baars and Stan Franklin. “Consciousness is Computational: The LIDA Model of Global Workspace Theory”. In: *Int. J. Mach. Conscious.* 01.01 (June 2009), pp. 23–32. ISSN: 1793-8430. DOI: `10.1142/S1793843009000050`.

[Bla02] Susan Blackmore. “There Is No Stream of Consciousness”. In: *J. Conscious. Stud.* 9.5-6 (2002), pp. 17–28. ISSN: 13558250.

[Bla03] Susan Blackmore. “Consciousness in Meme Machines”. In: *J. Conscious. Stud.* 10.4-5 (2003), pp. 19–30.

[Bla12] Susan Blackmore. “Could a machine be conscioius?” In: *Introd. to Conscious.* 2nd Editio. New York: Oxford University Press, 2012. Chap. 17, 270ff. ISBN: 978-0-19-973909-7.

[Blo02] Ned Block. “Some Concepts of Consciousness”. In: *Philos. Mind Class. Contemp. Readings* 18 (2002), pp. 206–219.

[Blo90] Ned Block. “Consciousness and accessibility”. In: *Behav. Brain Sci.* 13.04 (Dec. 1990), pp. 596–598. ISSN: 0140-525X. DOI: `10.1017/S0140525X00080316`.

[Bos98] Nick Bostrom. “How Long Before Superintelligence”. In: *Int. Jour. Futur. Stud.* 2 (1998).

[BS11] Adam B. Barrett and Anil K. Seth. “Practical Measures of Integrated Information for Time-Series Data”. In: *PLoS Comput. Biol.* 7.1 (Jan. 2011). Ed. by Olaf Sporns, e1001052. ISSN: 1553-7358. DOI: `10.1371/journal.pcbi.1001052`.

[Cer15] Michael A. Cerullo. “The Problem with Phi: A Critique of Integrated Information Theory”. In: *PLOS Comput. Biol.* 11.9 (Sept. 2015). Ed. by Konrad P. Kording, e1004286. ISSN: 1553-7358. DOI: `10.1371/journal.pcbi.1004286`.

[CF03] I D Couzin and N R Franks. “Self-organized lane formation and optimized traffic flow in army ants”. In: *Proc. R. Soc. B Biol. Sci.* 270.1511 (Jan. 2003), pp. 139–146. ISSN: 0962-8452. DOI: `10.1098/rspb.2002.2210`.

[Cha95] David J Chalmers. “Facing Up to the Problem of Consciousness”. In: *J. Conscious. Stud.* 2.3 (1995), pp. 1–23. DOI: `10.1093/acprof`.

[Cha97] David J Chalmers. *The Conscious Mind.* Oxford University Press, 1997. ISBN: 9780195117899.

[CM07] Antonio Chella and Riccardo Manzotti. “Artificial Intelligence and Consciousness”. In: *Impr. Acad.* (2007).

[Cri+04] Francis Crick et al. “Consciousness and Neurosurgery”. In: *Neurosurgery* 55.2 (Aug. 2004), pp. 273–282. ISSN: 1524-4040. DOI: `10.1227/01.NEU.0000129279.26534.76`.

[Edl+11] Jeffrey A Edlund et al. “Integrated Information Increases with Fitness in the Evolution of Animats”. In: *PLoS Comput. Biol.* 7.10 (Oct. 2011). Ed. by Lyle J. Graham, e1002236. ISSN: 1553-7358. DOI: `10.1371/journal.pcbi.1002236`.

[FKM98] Stan Franklin, Arpad Kelemen, and L. McCauley. “IDA: a cognitive agent architecture”. In: *SMC’98 Conf. Proceedings. 1998 IEEE Int. Conf. Syst. Man, Cybern. (Cat. No.98CH36218).* Vol. 3. IEEE, 1998, pp. 2646–2651. ISBN: 0-7803-4778-1. DOI: `10.1109/ICSMC.1998.725059`.

[FP06] Stan Franklin and F.G. Patterson. “The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent”. In: *Integr. Des. Process Technol.* (2006), pp. 1–8.

[Fra+12] Stan Franklin et al. “Global Workspace Theory, its LIDA model and the underlying neuroscience”. In: *Biol. Inspired Cogn. Archit.* 1 (July 2012), pp. 32–43. ISSN: 2212683X. DOI: `10.1016/j.bica.2012.04.001`.

[Fra00] Stan Franklin. “Deliberation and Voluntary Action in ’conscious’ Software Agents”. In: *Neural Netw. World* 10 (2000), pp. 505–521.

[Gam08] David Gamez. “Progress in machine consciousness”. In: *Conscious. Cogn.* 17.3 (Sept. 2008), pp. 887–910. ISSN: 10538100. DOI: `10.1016/j.concog.2007.04.005`.

[Gel09] Petros Gelepithis. “Outline of a new approach to the nature of mind”. In: (2009). URL: `http://cogprints.org/6571/`.

[Gel14] Petros a. M. Gelepithis. “A Novel Theory of Consciousness”. In: *Int. J. Mach. Conscious.* 06.02 (Dec. 2014), pp. 125–139. ISSN: 1793-8430. DOI: `10.1142/S1793843014400150`.

[GGT07] Simon Garnier, Jacques Gautrais, and Guy Theraulaz. “The biological principles of swarm intelligence”. In: *Swarm Intell.* 1.1 (Oct. 2007), pp. 3–31. ISSN: 1935-3812. DOI: `10.1007/s11721-007-0004-y`.

[GS12] Selvi Elif Gök and Erdinç Sayan. “A philosophical assessment of computational models of consciousness”. In: *Cogn. Syst. Res.* 17-18 (July 2012), pp. 49–62. ISSN: 13890417. DOI: `10.1016/j.cogsys.2011.11.001`.

[Hau+16] Andrew M. Haun et al. “Contents of Consciousness Investigated as Integrated Information in Direct Human Brain Recordings”. Feb. 2016.

[Hol03] O. Holland. *Machine Consciousness.* IMPRINT ACADEMIC, 2003. ISBN: 978-0907845249.

[Hol07] Owen Holland. “A Strongly Embodied Approach to Machine Consciousness”. In: *J. Conscious. Stud.* 14.7 (2007), pp. 97–110.

[HP14] Stuart Hameroff and Roger Penrose. “Consciousness in the universe”. In: *Phys. Life Rev.* 11.1 (Mar. 2014), pp. 39–78. ISSN: 15710645. DOI: `10.1016/j.plrev.2013.08.002`.

[INI10] Hiroyuki Ishiwata, Nasimul Noman, and Hitoshi Iba. “Emergence of Cooperation in a Bio-inspired Multi-agent System”. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics).* Vol. 6464 LNAI. 2010, pp. 364–374. ISBN: 3642174310. DOI: `10.1007/978-3-642-17432-2_37`.

[KJM15] Igor V. Karpov, Leif M. Johnson, and Risto Miikkulainen. "Evaluating team behaviors constructed with human-guided machine learning". In: *2015 IEEE Conf. Comput. Intell. Games.* IEEE, Aug. 2015, pp. 292–298. ISBN: 978-1-4799-8622-4. DOI: 10.1109/CIG.2015.7317946.

[Kli+15] P Christiaan Klink et al. "Theories and methods in the scientific study of consciousness". In: *Const. Phenomenal Conscious.* 2015, pp. 17–47. DOI: 10.1075/aicr.92.02kli.

[KMS09] Lukas König, Sanaz Mostaghim, and Hartmut Schmeck. "Decentralized evolution of robotic behavior using finite state machines". In: *Int. J. Intell. Comput. Cybern.* 2.4 (Nov. 2009). Ed. by Suranga Hettiarachchi, pp. 695–723. ISSN: 1756-378X. DOI: 10.1108/17563780911005845.

[KO16] Stephan Krohn and Dirk Ostwald. "Computing Integrated Information". In: (Oct. 2016), pp. 1–35. arXiv: 1610.03627.

[Koc+16] Christof Koch et al. "Neural correlates of consciousness: progress and problems". In: *Nat. Rev. Neurosci.* 17.5 (Apr. 2016), pp. 307–321. ISSN: 1471-003X. DOI: 10.1038/nrn.2016.22.

[KT07] Christof Koch and Naotsugu Tsuchiya. "Attention and consciousness: two distinct brain processes". In: *Trends Cogn. Sci.* 11.1 (Jan. 2007), pp. 16–22. ISSN: 13646613. DOI: 10.1016/j.tics.2006.10.012.

[Lam10] Victor a. F. Lamme. "How neuroscience will change our view on consciousness". In: *Cogn. Neurosci.* 1.3 (Aug. 2010), pp. 204–220. ISSN: 1758-8928. DOI: 10.1080/17588921003731586.

[Leg08] Shane Legg. "Machine Super Intelligence". PhD thesis. 2008.

[Lim+96] J.a. Lima et al. "Fitness function design for genetic algorithms in cost evaluation based problems". In: *Proc. IEEE Int. Conf. Evol. Comput.* October. IEEE, 1996, pp. 207–212. ISBN: 0-7803-2902-3. DOI: 10.1109/ICEC.1996.542362.

[Mag+14] Phil Maguire et al. "Is Consciousness Computable? Quantifying Integrated Information Using Algorithmic Information Theory". In: *arXiv Prepr. arXiv1405.0126* (May 2014), pp. 2615–2620. arXiv: 1405.0126.

[ME16] Alexander Maye and Andreas K Engel. "The Sensorimotor Account of Sensory Consciousness". In: *J. Conscious. Stud.* 23.5-6 (2016), pp. 177–202.

[MHA13] Lars Marstaller, Arend Hintze, and Christoph Adami. “The Evolution of Representation in Simple Cognitive Networks”. In: *Neural Comput.* 25.8 (Aug. 2013), pp. 2079–2107. ISSN: 0899-7667. DOI: 10.1162/NECO_a_00475. arXiv: 1206.5771.

[Mii+12] Risto Miikkulainen et al. “Multiagent Learning through Neuroevolution”. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics).* Vol. 7311 LNCS. 2012, pp. 24–46. ISBN: 9783642306860. DOI: 10.1007/978-3-642-30687-7_2.

[Nag74] Thomas Nagel. “What Is It Like to Be a Bat?” In: *Philos. Rev.* 83.4 (Oct. 1974), p. 435. ISSN: 00318108. DOI: 10.2307/2183914. arXiv: arXiv:1011.1669v3.

[OAT14] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. “From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0”. In: *PLoS Comput. Biol.* 10.5 (May 2014). Ed. by Olaf Sporns, e1003588. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003588.

[Oiz+16] Masafumi Oizumi et al. “Measuring Integrated Information from the Decoding Perspective”. In: *PLOS Comput. Biol.* 12.1 (Jan. 2016). Ed. by Daniel Polani, e1004654. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004654. arXiv: 1505.04368.

[Ols+12] Randal S Olson et al. “Predator confusion is sufficient to evolve swarming behavior”. In: *J. R. Soc. Interface* 10 (Sept. 2012), p. 20130305. ISSN: 1742-5662. DOI: 10.1098/rsif.2013.0305. arXiv: 1209.3330.

[Ols15] Randal S Olson. “Elucidating the Evolutionary Origins of Collective Animal Behavior”. PhD thesis. 2015.

[ON01] J. Kevin O’Regan and Alva Noë. “A sensorimotor account of vision and visual consciousness”. In: *Behav. Brain Sci.* 24.05 (Oct. 2001), pp. 939–973. ISSN: 0140-525X. DOI: 10.1017/S0140525X01000115.

[Pér+07] Óscar Pérez et al. “Comparison Between Genetic Algorithms and the Baum-Welch Algorithm in Learning HMMs for Human Activity Classification”. In: *Appl. Evol. Comput.* Vol. 4448. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 399–406. ISBN: 978-3-540-71804-8. DOI: 10.1007/978-3-540-71805-5_44.

[PRG16] André Luis O Paraense, Klaus Raizer, and Ricardo R. Gudwin. "A machine consciousness approach to urban traffic control". In: *Biol. Inspired Cogn. Archit.* 15 (Jan. 2016), pp. 61–73. ISSN: 2212683X. DOI: 10.1016/j.bica.2015.10.001.

[PRP16] Phil Maguire, Rebecca Maguire, and Philippe Moser. "Understanding Consciousness as Data Compression". In: *J. Cogn. Sci. (Seoul).* 17.1 (Mar. 2016), pp. 63–94. ISSN: 1598-2327. DOI: 10.17791/jcs.2016.17.1.63.

[Reg13] James A Reggia. "The rise of machine consciousness: Studying consciousness with computational models". In: *Neural Networks* 44 (Aug. 2013), pp. 112–131. ISSN: 08936080. DOI: 10.1016/j.neunet.2013.03.011.

[Rei+15] Chris R Reid et al. "Army ants dynamically adjust living bridges in response to a cost-benefit trade-off." In: *Proc. Natl. Acad. Sci. U. S. A.* 112.49 (2015), pp. 15113–8. ISSN: 1091-6490. DOI: 10.1073/pnas.1512241112.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. "The Earth Mover's Distance as a Metric for Image Retrieval". In: *Int. J. Comput. Vis.* 40.2 (2000), pp. 99–121. ISSN: 09205691. DOI: 10.1023/A:1026543900054.

[Sea00] John R Searle. "Consciousness". In: *Intellectica* 31 (2000), pp. 85–110.

[Sea90] J. R. Searle. *Is the brain's mind a computer program?* 1990. DOI: 10.1038/scientificamerican0190-26.

[Sea93] John R Searle. "The Problem of Consciousness". In: *Conscious. Cogn.* 2.4 (Dec. 1993), pp. 310–319. ISSN: 10538100. DOI: 10.1006/ccog.1993.1026.

[Sea97] John R. Searle. "Breaking the Hold: Silicon Brains, Conscious Robots, and Other Minds". In: *Nat. Conscious.* (1997), pp. 451–559.

[Skr03] David Skrbina. "Panpsychism as an Underlying Theme in Western Philosophy A Survey Paper". In: *J. Conscious. Stud.* 10.3 (Jan. 2003), pp. 4–46.

[SP11] Janusz A. Starzy and Dilip K. Prasad. "A Computational Model of Machine Consciousness". In: *Int. J. Mach. Conscious.* 03.02 (Dec. 2011), pp. 255–281. ISSN: 1793-8430. DOI: 10.1142/S1793843011000819.

[Teg15] Max Tegmark. "Consciousness as a state of matter". In: *Chaos, Solitons & Fractals* 76 (July 2015), pp. 238–270. ISSN: 09600779. DOI: `10.1016/j.chaos.2015.03.014`. arXiv: `1401.1219`.

[TK14] Giulio Tononi and Christof Koch. "Consciousness: Here, There but Not Everywhere". In: (May 2014). arXiv: `1405.7089`.

[TK15] Giulio Tononi and Christof Koch. "Consciousness: here, there and everywhere?" In: *Philos. Trans. R. Soc. B Biol. Sci.* 370.1668 (Mar. 2015), pp. 20140167–20140167. ISSN: 0962-8436. DOI: `10.1098/rstb.2014.0167`.

[Ton+16] Giulio Tononi et al. "Integrated information theory: from consciousness to its physical substrate." In: *Nat. Rev. Neurosci.* 17.7 (2016), pp. 450–61. ISSN: 1471-0048. DOI: `10.1038/nrn.2016.44`.

[Ton04] Giulio Tononi. "An information integration theory of consciousness". In: *BMC Neurosci.* 5.1 (2004), p. 42. ISSN: 14712202. DOI: `10.1186/1471-2202-5-42`.

[Ton12] Giulio Tononi. "Integrated information theory of consciousness: an updated account". In: (2012), pp. 290–326.

[Tur50] A. M. Turing. "Psychology and Philosophy". In: *Routledge Companion to Ninet. Century Philos.* Vol. LIX. 236. 1950, pp. 433–460.

[WT11] Josh L. Wilkerson and Daniel R. Tauritz. "A guide for fitness function design". In: *Proc. 13th Annu. Conf. companion Genet. Evol. Comput. - GECCO '11* (2011), p. 123. DOI: `10.1145/2001858.2001929`.

[XZL07] Jiyi Xiao, Lamei Zou, and Chuanqi Li. "Optimization of Hidden Markov Model by a Genetic Algorithm for Web Information Extraction". In: *Proc. Intell. Syst. Knowl. Eng.* (2007). DOI: `10.2991/iske.2007.48`.

[ZD14] Karolina Zawieska and Brian R. Duffy. "The Self in the Machine". In: *Pomiary Autom. Robot.* 18.2 (Feb. 2014), pp. 78–82. ISSN: 14279126. DOI: `10.14313/PAR_204/78`.

C List of Figures

1 Architecture of the GWT using the Theatre Metaphor 13

2 Integrated Information in mechanisms 24

3 Integrated Conceptual Information 25

4 Basic Workflow of a Evolutionary Algorithm 28

5 Architecture of a HMG 29

6 Architecture of a HMG 29

7 The Level Design 32

8 Workflow of one Generation in the Evolutionary Process 37

9 Abstract Process of the Evaluation Model 38

10 MABE functional overview 40

11 Basic Workflow of the Execution and Evaluation Process 41

12 Entity-Relationship Diagram 42

13 Multithreading is Used to Speedup the Overall Runtime 43

14 Experiments A 49

15 Experiments B 50

16 Wiring Diagrams of the Best Performing Genome of Animat α . 56

17 State With the Highest Φ^{Max} of the Best Genome of Animat α . 56

18 Collective Behavior of Animats Visualized 57

19 All LODs for Experiment Setting α III

20 All LODs for Experiment Setting β III

21 All LODs for Experiment Setting γ IV

22 All LODs for Experiment Setting δ IV

23 All LODs for Experiment Setting ϵ V

24 All LODs for Experiment Setting $\alpha * 0.75$ V

25 All LODs for Experiment Setting $\alpha * 0.50$ VI

26 All LODs for Experiment Setting ζ VI

27 All Heat Maps for the Best Animat in an Evolution for Animat αVII

28 All Heat Maps for the Best Animat in an Evolution for Animat βVII

29 All Heat Maps for the Best Animat in an Evolution for Animat γVIII

30 All Heat Maps for the Best Animat in an Evolution for Animat δVIII

31 All Heat Maps for the Best Animat in an Evolution for Animat ϵ IX

32 All Heat Maps for the Best Animat in an Evolution for Animat $\alpha * 0.75$ IX

33 All Heat Maps for the Best Animat in an Evolution for Animat $\alpha * 0.50$ X

34 All Heat Maps for the Best Animat in an Evolution for Animat ζ X

D List of Tables

1 Consciousness in IIT 20

2 Axioms and Postulates Explained 22

3 Animates Explained 33

4 Definition of the Mathematical Notation for the Fitness Function 34

5 Evaluation of the Top Performers A 53

6 Evaluation of the Top Performers B 54

7 Average Correlation Coefficients 55

E Glossary

AI Artificial Intelligence. 7, 11, 14, 16, 18, 24, 59

ANN Artificial Neural Network. 39

cei cause-effect-information. 23

ci cause-information. 23

CM connectivity matrix. 36

EA Evolutionary Algorithms. 27, 28, 39

ei effect-information. 23

EMD Earth Mover's Distance. 23

GA Genetic Algorithm. 26, 27, 36, 39, 40, 46, 58

GWT Global Workspace Theory. XVIII, 10–14, 20

HMG Hidden Markov Gate. XVIII, 28, 29, 46, 56, 58

HMM Hidden Markov Model. 39

IDA Intelligent Distributed Agent. 12, 13

IIT Integrated Information Theory. XIX, 1–6, 11, 14, 19–26, 28, 30, 31, 36, 40, 44, 45, 49, 50, 55, 59

LIDA Learning Intelligent Distributed Agent. 12, 14

LOD Line of Decent. 47, 51, 53, 54

MABE Modular Agent Based Evolver. 39–42, 46

MC Machine Consciousness. 1, 7, 14, 16, 23, 36

NCC Neural Correlates of Consciousness. 10, 19, 26

SEM Standard Error of the Mean. 47, 49–51

SMC Sensorimotor Contingency. 10, 11

TPM Transition Probability Matrix. 36, 39

WM Working Memory. 12, 13

Statutory Declaration

I assure that this thesis is a result of my personal work and that no other than the indicated aids have been used for its completion. Furthermore I assure that all quotations and statements that have been inferred literally or in a general manner from published or unpublished writings are marked as such. Beyond this I assure that the work has not been used, neither completely nor in parts, to pass any previous examination.

Neukirchen, 12.04.2017 *Dominik Fischer*