TOWARDS MOTIVATED AGENTS:

AN IMPLEMENTED BLUEPRINT OF A

COMPUTATIONAL MOTIVATION THEORY

IN MINECRAFT

by

Reinier Tromp

A document submitted in partial fulfillment of the requirements

for the degree of MSc Artificial Intelligence

at

Utrecht University

July 2019

Abstract: If we want to understand why intelligent organisms move, we need to account for motivation. Artificial Intelligence provides a suitable approach for investigating motivation. Likewise, motivation is crucial for reaching Artificial General Intelligence. However, motives are still lacking in AI-models today. To investigate the general requirements of such a computational theory of motivation, multiple motivation theories are examined and compared. This is used to evaluate a promising computational theory of motivation that is under development for the MicroPsi cognitive architecture. To further develop and test the hypotheses of motivation theory it is implemented in an agent in the virtual world Minecraft. This provides a basis for research on motivated agents in virtual worlds.

Acknowledgments

This thesis would not have been possible without the guidance of my professor Jan Broersen who kept me on track and helped me structuring my writings and my associative mind. I would like to thank my daily supervisor Joscha Bach for guiding me through the world of AI during our many hours of conversation. Special thanks to Chris Janssen for his supervision and confidence. Thanks to Dirk Meulenbelt for collaboration and discussion. And of course I owe my deepest gratitude to my parents, Rein and Wilma and my girlfriend Maartje who have always supported me.

Contents

Chapter 1

How the mind works

1.1 Motivation for Cognition

In a paper in 1967, Herbert Simon, one of the founders of AI seemed to be largely satisfied about the progress that had been made by his field in accounting for human cognitive performances. Yet, he proposes what he thinks is still generally missing in AI models:

> "Information-processing theories, however, have generally been silent on the integration of cognition with affect. Since in actual human behavior, motive and emotion are major influences on the course of cognitive behavior, a general theory of thinking and problem solving must incorporate such influences." (Simon 1967, p.29)

Since Simon wrote these words, the first of the two major influences on cognitive behavior has been seriously incorporated in AI research. Due to work of Aaron Sloman (1981) and Marvin Minsky (2007), the study of emotion received its place in the philosophical and scientific areas of AI. Especially when we look at the more engineering aim of AI, emotion has become salient. If we take for example the field of Human- Computer-Interaction we see that emotion detection and simulation has

become central for developing computer companions that aim to interact with people in meaningful ways without creating too much discomfort for the user (Boden 2016, p.72 - 77).

Unfortunately however, the second major influence on the course of cognitive behavior that Simon addressed, has not received much attention of information-processing theories. The study of motivation has until this day been largely absent from theories that try to account for human cognitive processes. This study will help changing that.

Yet before I motivate why motivation should be part of any theory of mind, let us first look at why accounting for human cognitive processes is at the core of AI research to begin with.

1.1.1 The original quest of artificial intelligence

For one who looks at the field of AI today, whether it are job openings, research groups, academic articles, master theses or newsletters, it is almost impossible not to think that work and research in AI is mostly about developing self learning algorithms of some sort that improve the accuracy of predictions or that it deals with designing programs for better problem solving in a particular task. However, if we give the field a second look we shall also discover a second - more philosophical and general approach that asks deep scientific questions about the nature of the human mind.

Similarly, Margaret A. Bodens (2016, p.2) distinguishes two aims of AI. One is *technological*: using computation to get things done that has utility and that doesn't need to have resemblances with the workings of the mind, also called "narrow AI". The other aim is *scientific*: using computation to help answer questions about living organisms - especially the human mind. This often gets refereed to as "Artificial General Intelligence" (AGI).

The second - scientific goal of AI is the underlying driving force of this thesis. It is centered around the question of how the mind works; the central question raised by John McCarthy, Marvin Minsky, Herbert Simon, Allen Newell and other early AI pioneers when they founded the field in the early 50's. Since then, the quest has been on to create artificial *humanlike* intelligence.

The general idea of Minsky (1986), Newell (1990) and others, was that psychology had stopped asking the right questions and relied on too narrow behavioristic methods that stopped them with making theories about the mind as a whole.[1] Whereas the ambition of early AI was to develop a unified theory of cognition that acknowledged the complexity of the human mind. Similar to physics, they believed that we can build theories expressed in a formal language to a very high degree of detail until we can simulate, run and test them. After all, computer science models have this additional criterion on top of needing to be capturing the regularities that you observe in the lab: they also need to work.

A working theory about the mind should therefore be a functional model. It must produce certain dynamics and reproduce the phenomenon that is under investigation. If it doesn't, we know that our theory is wrong. A good example is vision. If we have an idea of how stereoscopic vision works, we don't test it by taking a set of people in a lab and measure their timing and see if this is congruent with the timing predicted by our theory. What we do is, we implement our theory and we see if our working model is able to resolve images that are stereoscopic into a 3D-presentation. The key is that it is something that we consider to be structurally equivalent; that is able to capture the necessary regularities.

Herewith, the scientific aim of AI can be pictured as a creative version of what Daniel Dennet calls the "design stance"(Bach 2007, iv). In *The Intentional Stance* (1987) Dennet distinguishes tree stances: The physical stance largely looks at the

[1] see chapter 2

physical constitution of systems. The design stance operates on a more abstract layer and doesn't look at the physical constitution but at how things are constructed: how it functions and what the purposes ares. In the intentional stance we are concerned with phenomena as beliefs, desires and thinking. Explanations on this level are given in terms of mental states. I believe that AI is our best bet for getting to know how the mind works because AI takes this creative design stance. It doesn't only describe how the system is constructed, it is the position of an engineer. The goal is to build a functional model of the thing in question. That means that AI can put philosophy of mind to the test: it can proof its theories by building them.

We should not forget that this is not a new idea. Already in the eighteenth century, the mathematician Gotlieb Leibniz (1714, 17) performed a thought experiment:

> Imagine there were a machine whose structure produced thought, feeling, and perception; we can conceive of its being enlarged while maintaining the same relative proportions - among its parts - , so that we could walk into it as we can walk into a mill. Suppose we do walk into it; all we would find there are cogs and levers and so on pushing one another

Leibniz already described what would be the goal of what we now call artificial intelligence: trying to understand the mind by constructing a working model of it.

Leibniz himself however didn't believe the proposal. He concludes:

> [P]erception and what depends on it, is inexplicable in a mechanical way, that is, using figures and motions

Until this day, many scientists who investigate the workings of the mind agree with the "AI-scepticism" of Leibniz. People like John McDowell, John Searle, Ned Block, Roger Penrose, Noam Chomsky, the late Hillary Putnam - to name a few - all agree that AI's tenet as a method of understanding and superseding human intelligence is doomed to fail. Probably contrary to public belief, if we look at the field of AI today,

these scientists have indeed been right: the original idea of building a working model for the mind has regressed into a multitude of isolated domains like robotics, neural- or deep learning, computer vision, logic, semantic web, etc. Each with different research questions and methods and without the aim to create a unified theory of cognition or reverse engineering human intelligence.

So only if we look at Boden's *technological* aim of AI, it is correct to say that AI has been very successful. In almost all areas of life AI technologies are used, designed for uncountable specific tasks: vision, reasoning, learning, predicting, language, problem solving and so on. The presupposition of much of this work in AI is that learning about the subproblems, in isolation, contributes to answering the big question of general AI and if not - it contributes to technology in our modern society, whether that is autonomous driving, diagnosing diseases, improving e-commerce or playing Go.

If we look at the *scientific* aim however, we are far from reaching anything that integrates these technologies to model the human mind. Even in the domain of AGI that has put this aim at the center of its scientific efforts, it is admitted that progress for reaching general intelligence isn't at all impressive. In the words of Minsky (quoted in: Goertzel and Pennachin 2007, v):

> "Only a small community has concentrated on general intelligence. [...]. The bottom line is that we really haven't progressed too far toward a truly intelligent machine. We have collections of dumb specialists in small domains; the true majesty of general intelligence still awaits our attack.We have got to get back to the deepest questions of AI and general intelligence.."

With this study I want to get back to the deepest questions of AI and contribute to building a whole, functionalist architecture of the human mind. This requires a dissection of the cognitive system into parts, study one of them, make theories

about it, build and test it and in the end discover the relationships between the other different parts to integrate them into a coherent framework. I believe that if we want to move forward in finding out how minds work and therewith in AI in general, motivation is an essential and fundamental part to study, model and implement.

1.2 Two reasons for exploring a theory of motivation for Artificial Intelligence

There are two main reasons why we should look at a theory of motivation. The first is the one we already found in Herbert Simon: We want our (computational) theory of mind account for the behavior of humans, not the capabilities of computers (Simon 1967). The second reason goes the other way around: discuss the capability of (super) computers and how to program them, not the capabilities of humans.

1.2.1 Reason 1: Motivation could be at the core of behavior and can account for human generality and flexibility.

The most obvious reason for incorporating a theory of motivation in information processing theory is because we want our theory of mind answer the question: "what moves us?". Or, "Why do we get out of bed in the morning?" "Why did she do that?" Or, in more scientific terms: What causes behavior? And, why does behavior vary? Why does behavior start and once begun, why is it sustained over time? Why do we pursuit different goals at different times? One of the most interesting characteristics of the behavior of organisms is that it is spontaneously organized: it seems first energized and then directed. Unfortunately we cannot see the underlying mechanism that causes this behavior. We know for example that we have desires and beliefs, but we don't know what constitutes them. If we want to understand why people act in

the first place, we have to come up with theories that try to describe these underlying mechanisms and explain why and how behavior gets caused. Simpler put, we have to develop theories that explain why we get up in the morning, strive for power, explore the environment or try to get food on the table.

In his famous paper "Computing Machinery and Intelligence"(1950), Alan Turing explains how new machines could model the processes happening in the human mind that involve techniques that require intelligence. Since, the focus of AI has been largely on intelligence: the ability to solve hard problems (Minsky 1986, p.71) or more concrete: the pursuit of goals in the face of obstacles (Pinker 1997).[2]

But now imagine that we put such an intelligent machine into the world. Instead of asking how it pursues goals, we can ask ourselves: why does it pursue goals at all? How does it create and select them? Intelligence may be seen as problem solving given a set of goals; a conventional computer program that executes a line of written code until it reaches the end. But it seems highly unlikely that organisms are restricted to those fixed set of tasks and goals. I believe that human generality and flexibility stem largely from creating, prioritizing and identifying those goals. Surely, intelligent organisms need a flexible method of control. Following Simon (1967), I believe that in an open environment with infinite new situations and events it is impossible to specify the complete set of goals a priori. Therefore an artificial system that tries to model the human mind, that tries to explain how the human mind is able to do what it does, it needs to account for the autonomy it.

[2]Minsky admits that the concept of intelligence is hard to define. Instead of trying to say what it means, it is better to try to explain how we use it, that is for problems that involve processes that we don't yet understand

1.2.2 Reason 2: Motivation could help answering questions of concern around "superintelligence".

There is also a technological and practical reason for accounting for motivation in information theories. This concerns the ethics of AI, especially the issue of design and controllability. The question here is how we can build intelligent systems (humanlike or not) that are save and good. One of the biggest concerns nowadays in AI is the danger of what is called "superintelligence". At the moment, machines are far inferior to humans in general intelligence, but as Nick Bostrom argued in his highly influential book *Superintelligence* (2016) we have to take into account the possibility that machines once will transcend us in general intelligence.[3] Bostrom warns specifically for the case that superintelligence will have the possibility to shape the future with its goals (2016, chapter 6 and 7). Later on, I will discuss why the focus on goals is typical for AI research but let us first take a closer look at the implications of Bostrom's argument.

Firstly, let us for the sake of the argument, accept the idea that superintelligence is indeed a possibility. Then our next two questions should be:

1. When it arrives, can we control it and if yes, how?

2. What does it want?

Or as phrased more popular in the recent film Arrival (2016):

> **Colonel Weber**: Priority one: what do they want and where are they from?
>
> **Ian Donnelly**: And beyond that, how did they get here? Are they capable of faster-than-light travel? We prepared a list of questions to,

[3]Bostrom concern gets support from many influential scientists, writers, thinkers, and businessleaders inside and outside the field of AI, like: Stephan Hawking, Elon Musk, Bill Gates, Peter Norvg, Joscha Bach, Stephen Fry, Ray Kurzweil, Sam Harris, Ben Goerzel and many others. This doesn't make its true, but shows a demand for answering questions of concern about the development of AI

you know, go over, starting with a series of just a handshake binary sequences...

Dr. Louise Banks: How about we just talk to them before we start throwing math problems at them?

What we should expect according to Bostrom is something intelligent *and* non-human. We shouldn't expect that becoming more intelligent means the same as becoming more human. Bostrom calls this the "anthropomorphizing of AI": projecting human values on a non-human entity. This means that apart from intelligence, we need to take a look at its core inner workings. Does its inner machinery produce the same (ethical) goals as us? In other words: what motivates it? And do these motives prevent them from doing us harm?

Humans seem to feel high level emotions like empathy because we have evolved to feel them (Pinker 1997). That is, in our inner workings we have something that feels pain or pleasure when someone similar does. We can put ourselves in someone else's shoes. But there is no reason why we should expect a superintelligence to feel the same thing. More intelligent doesn't mean more human or more empathic. This is the case for humans, where there is no correlation between intelligence and morality or empathy. That is, intelligence measured in IQ. A superintelligent agent could very well be totally emotionless or amoral. Unless we program it to be.

In the interesting episode "More or less Human" produced by the science and technology podcast Radiolab, the radio makers performed an experiment with what they call "an emotional Turing Test". A group of children between 6 and 8 years old were asked to hold 3 objects upside down: A Barby, a Furby - a furry doll that starts crying when turned on its head - and a hamster, called Jirby. For instructions, the children were asked while they got handed out each of the three objects: "how long can you keep it upside down before you yourself feel uncomfortable". Then, the instructors timed the estimated time each children hold each object on its head. The

longest "upside down time" was measured for Barby with an average of 5 minutes, until the participants got lazy arms practically. After Barby followed Furby with an average of 60 seconds. The shortest upside down time was for the hamster Jirby with an average time of 8 seconds.

When the children were asked why they experienced difficulty with holding Furby upside down compared to Barby they mentioned that they were not sure if Furby "could feel pain". The children - the radio producers concluded - clearly had mixed feelings about Furby's affective state. Or, as one kid articulated it: "two thoughts: the first is: 'it's a toy for crying out loud!' And the other is: 'it makes me feel guilty!'"

Similar examples are provided by the Milgram-style experiments with avatars (Misselhorn 2009). Like the Radiolab experiment, the results of these sort of experiments show that in spite of the fact that participants know for sure that objects are inanimated, once they get confronted with the emotional outcry of them, they respond with a higher level of empathy - almost up to the level of dealing with real human beings. Clearly, we feel more empathy for animated objects than for inanimated objects, but interesting enough we also feel more empathy for something that *behaves like* an animated object.

Therefore, according to Bostrom (2016, chapter 7) we need to ask: what motivates an AI system? The warning of antropomorphization he says, should be extended to motivation of AI as well. From a random intelligent alien it would not be highly surprising that its motivation is related to items like food, water, air, body temperature, energy expenditure, threat of body injury, disease or sex. If it is also a social species, it would probably also have motives related to competition and cooperation.

A superintelligent agent by contrast, need not care intrinsically about any of those feelings. It is even more likely to have an AI whose sole final goal is simple, like maximizing the total number of paper clips, than it is to have an AI with human values and dispositions. According to Bostrom, intelligence level and *final goals* are

orthogonal: any level of intelligence can be combined with any final goal. AI is whatever we programmed its topmost goal to be and it's utility function may not be perfectly aligned with human values, which are very hard to pin down.

The key move Bostrom makes is that he combines his orhtogonal thesis with what he calls *instrumental convergence* (p.131). The latter means that when you try to achieve a long-reaching goal, you often aim for several subgoals along the way. If you want to produce a paper clip, you first need to get out of bed, make breakfast and fix your flat tire. We humans, Bostrom argues, often let our final goals drift because of *continuously self-discovery* and *changing needs* (p.134). We tend to seek and acquire resources sufficient to meet our biologically needs and often go beyond that to satisfy other needs, like convenience, social status, friends and influence through wealth and consumption or altruistic behavior (p.137).

Whether or not a superintelligent agent will use all the earth resources for the benefit of scaling depends on the final goals it has, and according to Bostrom because of the combination of the orthogonal thesis and the instrumental convergence thesis, that story is not going to end well for human beings. Any intelligent system will try to succeed in its assigned task and therefore will secure its own survival as long as necessary and acquire the resources that it needs. It will do whatever it costs to pursuit it's topmost goal.

Therefore, the natural question for a superintelligent agent seems to be: where do his goals come from? Bostrom's gives us a goal-oriented answer - without giving it much explanation. Depending on the final goal, he says, we will know its sub-goals and therefore know what its motivation is.

The major difficulty however I have with this top-down approach, is that for the programmer it is almost impossible to foresee all the possible values of the unconstrained variables of the function that the AI system is optimizing for. This difficulty is called "the alignment problem" (Hadfield-Menell et al. 2016). We should

be familiar with the alignment problem because it has been a returning theme in both classic and modern literature like the fable of *King Midas,* Goethe's poem *The Sorcerer's Apprentice* or J.K Rowling's *Tale of Three Brothers* in *Harry Potter and The Deadly Hallows*[4]. The gist of these fables is that when you get what you wish, you do not necessary get what you want. The apprentice in *The Sorcerer's Apprentice* gets tired of fetching water and enchants a broomstick to do the work for him. He soon finds out that the broomstick's utility function is optimized for bringing water to a reservoir whether or not the whole workshop gets flooded along the way (the unconstrained variable). To make things worse, the enchanted broomstick is missing an off-switch. It is not the case that the broomstick is driven by its survival instinct, nor is it trying to dominate its master. It only tries to achieve its topmost goal: bringing water from the fountain to the reservoir and meanwhile - as an unfortunate side effect - it almost drowns its master.

All things considered, the enchanted broomstick follows Bostrom's orhtogonal thesis without applying instrumental convergence. Similar to the apprentice, a programmer of a superintelligent agent can not foresee or understand all the consequences and subgoals that follow from the final goal of its program and therefore doesn't know if these subgoals and corresponding values are aligned with his or her purposes. In response, prominent AI-researchers call for research in AI that is "provably aligned" with human values to help ensuring that the future societal impact of AI is beneficial (Russell et al. 2015). Bostrom's top-down solution to this problem in which the programmer investigates the final goals conscientiously, has a hight risk of falling victim to the same fate as the apprentice in *The Sorcerer's Apprentice*.

So instead of discovering the motivation and subgoals of an agent by deducing its final goal, a solution for the orthogonal thesis problem could be looking at motivation itself. This turns Bostrom's solution on its head. Instead of programming its final

[4]See the beautifully animated version of this tale as part of the movie here: https://www.youtube.com/watch?v=aJSh1zkPEvc

goal and from there on align the agent values with human values[5], we program its motivation and from there on align its needs and drives, subgoals and final goals from the bottom-up. We let it create, prioritize and identify its own goals, within the range of possible needs that we endowed it with. It is therefore worthwhile, as I will argue in the chapter to follow, that we shouldn't occupy ourself with coding pre-defined goals but to see whether we can design a motivational system that comes up with desired goals in the first place.

1.3 Research questions

Motivation can be regarded as an essential part of understanding how the mind works. If we ask ourselves what makes us move? Or what happens if we put an intelligent agent in the world, we have to answer why we behave in the first place. An implemented theory of motivation is currently lacking in AI-frameworks that aim to understand intelligence. Therefore I want to investigate how a model of motivation would look like and see if this can be implemented. Therewith this thesis contributes to the scientific aim of AI and the original quest of AI to build humanlike intelligence or at least regard humanlike intelligence as inspirational source.

At the same time I believe that building a functional model of the human mind can inform research and applications with a technological aim. If we want to build responsible AI it is worthwhile to not only look at how we can derive intentions from its topmost goals. If we know the purpose of a tree (survival, produce offspring) it is - maybe possible - but still hard to derive how it behaves. Instead, if we look at what a tree needs (water, light, carbon dioxide and so on) this could provide us with a better model for predicting its manner of conducting itself and furthermore, how to manipulate it.

[5]It is questionable whether human values are desirable for AI. AI that doesn't harm humans is something different than AI that behaves humanlike.

This brings me to the following research questions:

1. What are the general requirements for a theory of motivation?

First I will answer how a model of motivation would look like. I don't want to reinvent the wheel here. Motivation has been a well-studied subject, especially during the first half of the twentieth century. I will investigate the most prominent theories, weight and evaluate their arguments and find the common ground between. This will give me the theoretical requirements for a model of motivation wherewith I can evaluate and build a computational model.

2. How does the computational model of motivation as developed for MicroPsi meet these requirements?

MicroPsi is a cognitive architecture[6] that claims to have motivation at the core of its framework. The architecture is still much work in progress, even more its motivational theory. By looking at Micro Psi's motivational model, its theoretical assumptions, properties and current configurations I will evaluate whether it satisfies the requirements of a motivational theory that resembles human cognition.

3. Can I implement a computational model of motivation in an agent that is situated in a virtual world and run it?

If we want to see if and how the motivational model truly functions and if we want to improve it with the findings done in the former questions, we have to run it. Because the ultimate goal of AI is understanding minds, and because minds are situated in the world, I have to implement the motivational model into an agent that is situated in the world.

[6]See chapter 3 for a definition

Shorty, with this thesis I want to contribute to our knowledge of how minds works and how they get into being. Therefore I will investigate of what I hold as the core of all living organism and where such an endeavor could start: motivation. For this I take Dennet's creative design stance. I will look at how a computational theory of motivation should look like based on existing theories of the past. Thereafter I investigate an ongoing effort of building a motivational model and review its strengths and weaknesses. Since the questions that I ask is what happens if we put an organism into the world, I will do exactly that: I will put a motivational agent in a world to see how it behaves, to provide further AI research with a workable blueprint. With this we can gain insight in how the mind works and investigate what is at the core of human cognition.

Chapter 2

Motivational Theory

If we want an information processing theory accounting for motivation, we need to do the following:

First we need to investigate existing theories of motivation to answer the question how a general theory of motivation would look like. Especially in the first half of the twentieth century attempts have been made to develop a grand theory of motivation that tried to explain how behavior gets energized and directed. Based on the obtained insights, we can design and build a computational implementation of motivation by taking a creative design stance: working on a functional model of the thing in question and see if it works by implementing it in a virtual world. In this chapter we will look at these existing theories of motivation.

2.1 The origins: Freud's Theory of Drives

2.1.1 Instinct theory

One of the first theories that provided us with the idea that animals are fundamentally motivated was Charles Darwin's theory of natural selection (Darwin 2011). Darwin showed that all species are motivated for survival and reproduction. Humans, just

as animals are equipped with "instincts": ways to do or make things, that are genetically transmittable (Dennett 2018, p.234). With this instinct theory (adopted and developed by William Wundt) Darwin could explain what philosophers before could not: where the motivational force of behavior comes from in the first place. Instincts are, he said, for both humans as animals the main drive behind the distribution of available resources and the explanation for our hardwired adaptive behavior (Reeve 2009, p.28).

Darwin's key insight for behavior was given by the idea that (animal) behavior seemed as unlearned, inherited and automated. Throw a ball to a pup and it will chase it like a predator. Do the same to a bird and it will try to dodge it. Baby turtles walk to the sea, baby mouth's make sucking movements and just as foals walk immediately after birth, they flee when they smell danger. These everyday acts are explainable: evolution is the force behind the genetically endowment of animals that gives rise to certain instincts that are the main forces behind all behavior.

One of the problems with instinct theory however- especially for scientific purposes - is that the logic underlying it seems circular: the cause (the instinct) explains the behavior but the only thing we can observe is the behavior itself, therefore we can only conclude from the behavior that the instinct exists. There seems to be no independent way examining both phenomena. Another problem with this theory of motivation is that it doesn't explain that organisms may exhibit different levels of motivation such as strive for power and social bonding. It therefore isn't powerful enough to explain the differences between different individuals.

2.1.2 Triebe und Triebschicksale

Influenced but unsatisfied with the state of instinct theory and also influenced by the insights of a new scientific field called physiology, Sigmund Freud (1915b, c) developed the first grand motivation theory: his theory of drives. He describes the

fundamentals in a paper called "*Triebe und Triebschicksale*" and called drive theory "the most important and darkest subject of psychological research"(8:192).[1]

First let us have a look at what Freud means by drive. His own definition of the term is never very clear and distinct and neither are the associated terms. He presents drive as a borderline case between body and mind - in Freud's own terms: between the psychical and the somatical. Drives are "motivational forces". They must be understood as representations from stimuli that are derived from the internal of the body that reach the soul (1915*c,* 7:27).

According to Freud, an organism consists of a nervous system that translates the organism's *needs* into the drives, that can also translate into *wishes* and *desires.* Freud makes it very clear that drives and needs should be strictly distinguished. The needs are internal to the organism, they are an "instinctual stimulus" (1915*c*, 7:24). Needs are therefore lower in hierarchy: they subsume to drives but they are not the solely determinants. Drives are affected by the needs but there are other - external - stimuli that act on them, like bright light that falls on the eyes and makes them close. These external stimuli are avoidable or can be sit out. Needs however are a constant force. They can cumulate and they require some form of action for relieve called *satisfaction.* Satisfaction can only be accomplished by an effective change in the internal source of the need, therefore the aim of the drive is to seek satisfaction by responding to the needs and reducing the *tension* that they give. This tension naturally increases if needs are neglected and stabilizing them is at the core of behavior. Reducing this

[1]In translation the German concept "*Trieb*" is consistently translated as "instinct". For example the paper's title has become: "Instincts and their Vicissitudes". However the concept "Instinct" is exactly what Freud tries to reconstruct. The German word for instinct is simply "Instinkt" and doesn't appear once in the original. *Trieb* however, is better translated as "drive", understood as some inner desire or force. The same holds for the translation of *schicksale* into "vicissitudes". Vicissitude is a descendant of the Latin noun *vicis*, meaning "change" or "alternation". The German concept *Schicksale* however, means "fate" or "destiny" and is used to refer to some higher power that influences ones life. Therefore *Triebschicksale* should be understood as something like a drive system that influences onces fate or destiny in life. We will see that understanding these concepts this way helps clarifying the core of Freud's theory of drives.

tension and regulating the psychical process, is what Freud later famously calls the *lustprinciple* (8:165).

The lustprincple and the object

With the lustprinciple we get into more psychoanalytical waters and that is beyond the scope of our concern. What is interesting however, is that Freud says that the lustprinciple chooses a direction based on inducing *pleasure* and avoiding *displeasure.* Here, Freud seems to make the first step to an account of learning. As we will see, one of the main criticisms of Freud's drive theory is that it leaves out any notions of reinforcement. Reinforcement would become one of the main concepts in early psychology literature and was the central pivot of any learning theory. Unfortunately, Freud never really incorporated nor generalized the characteristics of the lustprinciple into a theory of learning.

The directional goal of the lustprinciple is based on a chosen *object.* Without that object, a drive cannot accomplish its aim and achieve satisfaction. Compare the lustprinciple with a ship that gets moved up and down by the waves and gets shifted to the left and the right by the wind. An experienced helmsman knows that he has to orient himself on a fixed spot on the horizon to keep the ship on course. In Freudian terms, we can call this spot the object. That object can be anything; a lighthouse, a distinctive piece of the cliffs or a bright star. It is not important what the object is, as long as assigning and prolonging it keeps the subject on course to satisfy the needs and therewith reduces the tension. One object can serve multiple needs at the same time. Freud calls this *drive entanglement* (7:28). When the drive connects to the object very fervently however, he calls it *fixation* - one of his famous pathologies.

Freud admits that the origins of the drives fall outside the domain of psychology. Therefore, if we want to investigate these origins, we need a method that can somehow give insight to the inner mechanism at work. Freud sees no way than to investigate

the drives' goal. This opens the door to therapy: by investigating what someone drives and goals are and what motivates him or her; by performing psychoanalysis, we can get an idea of the underlying mechanism at work.

Kind of drives

But what different drives do humans have? Freud admits that classifying them is hypothetical:

> "it is doubtful whether any decisive pointers for the differentiation and classification of the instincts (drives, RT) can be arrived at on the basis of working over the psychological material." (7:30)

He hopes that other scientific domains, like biology can help with this effort but for now we have to treat the distinction between specific drives - like the "sexual drive" and the "drive for self-preservation" as helping constructs and shouldn't be taken too seriously.

2.1.3 Value and criticism of Freud's theory of drives

Freud's drive theory became very influential in the first half of the twentieth century. It was the first theory that combined Darwin's instinct theory with experience. It made a distinction between inner environments (needs) and outer environments (external stimuli) and showed how they influence each other. Furthermore, it could account for psychological phenomenons and mental illness like frustration (dissatisfaction of needs), fixation (assigning needs too fervently to a goal) and so on.

Despite its originality and explanatory power, Freud's theory of drives suffers from several weaknesses. First, it doesn't account for learning, something that started to receive much attention in the thirties and forties of the twentieth century. Freud speaks about pleasure and displeasure that affects our behavior, but he doesn't tell us what the effects are for future deeds or drives.

Secondly, Freud based his theory solely on case studies of disturbed individuals instead of on experimental results from representative samples. This of course limits the generalization of its claims severely.

Thirdly and connected to the former point, parts of the drive theory were not scientifically testable. Freud admitted himself already that we cannot investigate the drives directly. His solution was to look at the drive's aims: the part that seeks to satisfy the needs. This is oddly enough very similar to Bostrom's solution for instrumental convergence and his orhtogonal thesis: the idea that an AI-agent will always utilize its topmost goal and along the way aims for several sub-goals - which may be unknown to us.

Freud's method of choice is therapy. The general goal of Freudian therapy, is to bring repressed thoughts and feelings from the unconsciousness or subconsciousness into the consciousness in order to free the patient from suffering distorted emotions. What Freud seems to do in his analysis of drives is to turn this on its head: by analyzing his wishes, desires and dream accounts, we can draw conclusions about the underlying mechanism of the man. The only part we have access to is this consciousness. But here we see that Freud falls into the same trap as Darwin did with his instinct theory: the cause (the drive) explains the behavior but the only thing we can observe is the behavior, therefore we can only conclude from the behavior that the drive exists. Freud and his followers worked a lifetime on this problem of circularity, developing specific therapies that could shed a light on what lies in the unconsciousness and give the idea scientific recognition. Until this day, Freud's method of psychoanalysis for discovering the inner workings of the mind is highly contestable in the sciences that favors more limited but empirically tested ideas. In clinical psychology however it still functions as one of the main pillars for treatment yet under the name of psychodynamics.[2]

[2]The others pillars are (cognitive) behavioral therapy, system therapy and client focused psychotherapy

Despite its weaknesses, the theory of drives as a model of the mind equipped the sciences with a brand new framework. It provided us with a theory about what lies in the unconsciousness that is mainly responsible for our behavior. Especially the theory's focus on internal dynamics in combination with external events was a major contribution on Darwin's instinct theory. One could argue that once the aforementioned weaknesses are solved - we'll have a solid basis for a theory of motivation.

2.2 Hull's Drive Reduction Theory

A theory that closely resembles that of Freud, but deals with much of its weaknesses is the drive reduction theory developed in the forties by the American psychologist Clark Hull (1943). Hull follows the basics of Freud's model almost entirely.[3] Similar to Freud, Hull argues that all behavioral motivation comes from the pleasure of meeting a biological need. A need is the biological requirement of the organism and humans try to reduce the tension that results from not satisfying the need. We find the same ideas:

1. Drive energizes behavior
2. Drive has its origin in somatic needs
3. The function of behavior is to serve the needs

Hull calls the underlying principle that keeps the organism into motion "homeostasis". Homeostasis is the principle by which a thermostat operates: when the current temperature is lower that the target temperature, it switches on. When the current temperature is equal or higher than the target temperature, it switches off or turns on the air conditioning. In the same way the body, when it lacks a certain substance, it develops hunger for it and when it has enough it stops.

[3]Incidentally without giving the master much credit: In *Principles of Behavior* Freud is exactly never referenced

Recall that the main weaknesses of Freud's drive theory were its vague notion of learning and its weak scientific methods. These are exactly the two point where Hull´s drive reduction theory improves upon Freud's.

In drive reduction theory, drive doesn't *direct* behavior. Instead, behavior is directed by *habits* and habits come from learning. The forming of habits is explained by involving the central idea of behaviorism into the model: reinforcement. Reinforcement refers to anything that increases the probability of a response occurring in the future. It follows that every response that successfully reduces the tension by satisfying the needs, reinforces the forming of habits. This is what we can call learning. Hull developed the following formula (Hull 1943, p.403):

$$ {}_sE_r =_s H_r \times D $$

The variable E stands for "excitatory potential" that we must understand as a form of behavioral "energy". The subscript s refers to the stimulus and the r to the response. In combination ${}_sE_r$ refers to the energy of a response in the presence of a particular stimulus. Thus, the behavioral strength is the strength between a situation and the need-fulfilling response. ${}_sH_r$ refers to the habit strength, and D is the drive. Note the multiplication sign: behavior only occurs when the habit strength *and* the drive are nonzero (Hull 1943, p. 404-409).

Both the habit and the drive are internal processes. Later, Hull updated the model for accounting for external influences on behavior. When you are working on something and someone interrupts you by asking you a question, it is likely that you stop working and start talking, due to the stronger (external) incentive of the interruption. By incorporating external factors in its model, drive reduction theory could account for the complex dynamics of an organism in relation to its environment.

2.2.1 Value and criticism on reduction theory

The strength of Hull's model and one of the main reasons for its popularity at the time are its predictive properties that made it very suitable for laboratory experiments. One of the most popular experiments - already described by Hull himself - were maze experiments with rats under the condition of food deprivation (Hull 1943, p. 226 - 256). When you deprive a rat of food, you expect his food drive to increase over time until it becomes the dominant force of behavior, resulting in food searching. This was a breakthrough for drive theory. Hull showed that drive was predictable by manipulating the environmental conditions and changing the internal state of the organism. This experimental approach in combination with explanatory modeling started the beginning of a scientific study to motivation with less emphasis on biology and more emphasis on learning and experience (Reeve 2009).

However, not all human behavior can to be explained by the dynamic of drive reduction. Take behavior as curious exploration, investigation, manipulation, vigorous play or other spontaneous activities. One wonders what the necessary nutriments are for getting launched in a roller coaster, taking drugs at a party or climbing a dangerous mountaintop. Or take for instance a disorder as anorexia nervosa where food deprivation serves the strong desire to be thin. What somatic need is served here? The more researchers learned about specific cases, the more they concluded that one grand theory based on physiological drives could not explain the complexity of human behavior.

2.3 Maslow's Hierarchy of Needs

Probably the most famous of all motivational researchers is Abraham Maslow who made the attempt to formulate a positive theory of motivation that could account for more psychological behavior. In line with Clark Hull - an not coincidently in the

same time span - Malsow's theory of motivation tried to combine the clinical and observational as well as the experimental.

In his paper "A theory of human motivation"(1943), Maslow takes the same basic principle as Hull and Freud for behavior by saying that our behavior is driven by unmet needs. He agrees with Freud that a theory of motivation integrates "the wholeness of the organism" and that it should stress "a more central place for unconscious than for conscious motivations"(1943, p.370). Maslow however stresses that not the somatic drives are at the centering point. Like Freud, he states that drives are a borderline case between the physical and the somatical and that an act can serve multiple needs. Motivational behavior must therefore be understood as "a channel through which many basic needs may be simultaneously expressed or satisfied". Remarkably, Maslow agrees with Freud when he says that a list of drives will get us nowhere; he shows serious caution about any classification (1943, p.371). Especially this last point is noteworthy because his theory became famous for its classification with the strong image of a pyramid.

The pyramid is an imagination of Maslow's "hierarchy of needs". Human needs arrange themselves in hierarchies of "pre-potency" (1943, p.372). This is a big difference with the models of Freud and Hull. It means that you are urged to satisfy the basic needs first before you worry about other - higher order - needs. First you will try to satisfy the physiological needs, like "hunger", before you can think of things like writing a thesis. It follows that the higher order needs are depended on to lower order needs but, opposed to popular conception, it doesn't follow necessarily that the value of the basic needs should be at 100% before higher order needs are calling. When a basic need is *more or less* satisfied, Maslow writes, it will go away and our activities will be directed at meeting the needs higher up in the pyramid.

Maslow's theory of motivation differs in several other aspects from Freud and Hull as well. First of all, Maslow often speaks about "goals". He says that a theory of

motivation "should stress and center itself upon ultimate or basic goals rather than partial or superficial ones, upon ends rather than means to these ends"(1943, p.370). The difference between goals, needs, drives and motives are never very clear and distinct in motivation theory but what Maslow means with "goal" seems to be what Freud means with "object". For Freud the object is something that gets assigned to a need and determines the direction of behavior. For Maslow goals are the basic principle of the classification of needs that gives behavior its direction. It seems that for both Freud and Maslow the direction of behavior is given by the purpose of the organism that is specified in a goal. Motivation theory looks at the "functions, effects, purposes or goals of the behavior", not at the instigation he says (1943, p.392). We will see that this incorporation and special treatment of goals gets important of we look at the combination of AI and motivation theory.

The emphasize on goals has a practical reason as well. Just as Freud and Hull, Maslow admits that we cannot know the inner states of organism from the outside. Therefore our theory must acknowledge that we have to rely on those things that we can discover at the surface of the mind, like goals and effects. Freud's method is therapy, Hull derives the inner nature of man from the results of stimulus-response experimentation. Maslow's method is qualitative biographical analysis. He created a sample of 18 persons of whom he thought where self-actualized and looked at their common properties.

During his career, Malsow optimized his theory of motivation. However, he never optimized his method, that was vulnerable to subjectivity, bias and lacked an experimental foundation. In the preface of the third edition of his *Motivation and Personality* (1987, p.xii) Maslow admits:

> It is fair to say that this theory has been quite successful in a clinical, social and personological way, but not in a laboratory and experimental way. It has fitted very well with the personal experience of most people,

> and has often given them a structured theory that has helped them to make better sense of their inner lives. It seems for most people to have a direct, personal, subjective plausibility. And yet it still lacks experimental verification and support. I have not yet been able to think of a good way to put it to the test in the laboratory.

Despite its popularity, due to its methodological weaknesses his theory has a sway of folk-psychology around it. Especially because Maslow claims that needs are universal, his method clearly lacks the power to support it. A theory that tries to account for motivation should at its foundation apply to all individuals in all cultures. With a method that is based on biographical analysis, that seems impossible to proof.

2.4 Modern Motivational Theories

2.4.1 From grand theories to mini theories

Drive theory had great ambitions: it tried to explain how the behavior of organisms have both energy and direction. Motivation it says, is at the core of it: we humans are motivational beings. People like Freud, Hull and Maslow attributed great value to our understanding of how the mind works, because they dared to ask the big questions and tried to answer them. However, modern theory of human motivation have found several weaknesses in the grand theories and have come up with a multitude of additions and variation. This has lead to a multitude of "mini-theories of motivation" (Reeve 2009, p.35-40). Motivation psychologists were gaining so much new information that they preferred to create their own theories, explaining a specific part of motivation instead of building on a larger framework that explained the full range of motivation. A multitude of mini-theories of motivation got developed like achievement motivation theory, attributional theory of achievement motivation, cognitive dissonance theory,

expectancy ×value theory, self schema's, intrinsic motivation etc. All of them tried to explain a part of human motivation, instead of all of it.

Another important development that shaped motivational theories was the "cognitive revolution" in the sciences. The drive theories of Freud and Hull were grounded in biology and physiology rather than cognition - a common focus at the time. Especially the drive reduction theory was pretty behavioristic in the sense that it focused on stimulus response experimentations. In the behaviorist circles however, Hull was a still a mild behaviorist since he acknowledged the existence of inner states. Many of his colleagues, among whom the influential psychologist B.F. Skinner, argued that science would never advance by looking at these "internal states" like will and desires. Skinner wrote: "[H]uman behavior is a function of conditions, environmental or genetic. People should not object when a scientific analysis traces their behavior to external conditions" (Skinner 2002). Motivation in the strong behavioristic tradition was fully attributed to the environment and therefore to learning.

The cognitive revolution was a strong reaction to this behaviorism (Chomsky 1971) and had its effect on motivation theory with a new focus on thoughts, judgments and beliefs. These ideas however can already be found in the grand theories of motivation. Especially Maslow's theory showed much flexibility for embedding the cognitive concepts (e.g. expectations and goals) in favor of the biological and physiological concepts (the basic needs) that got deemphasized.

The same happened to methodology. The ease with which behaviorist generalized from rats to humans was already problematized in Maslow's early work in which he noted that "a motivation theory should be human-centered rather than animal centered"(Maslow 1943, p.371) and claimed that testing animals under stress of food deprivation, social isolation and outside their natural habitat is not telling you much about motivation in human beings. He also said that although organismic states are

to be understood as motivated and motivating, motivation theory should not be seen as equal to behavior theory. Motivation is just one class of determinants of behavior.

The cognitive turn gave a new position to motivation in the model of the human mind. Motivation is still at the center of it but behavior is just one of the many things that gets influenced. Motivation causes activity and activity has many forms, like dreams, aspirations, thoughts, feelings and indeed behavior. Behavior in itself also contains multiple complex forms like goal directness, initiation, termination, change and persistence that all need their own theories.

Another more practical reason for the rise of mini-theories at the cost of the grand theories of motivation is the growing importance of clinical psychology. Instead of developing a theory that tries to comprehend the workings of the human mind, the focus became solving motivational problems for the workforce, the educational system or for self management. This leads to different questions like "what is the role of inner motivation on learning" or "what is the role of reward in undermining interest" and "how does a teacher's praise affects students motivation?" (Pintrich 2003). What emerged where studies of social motivation, developmental motivation, motivation underlying losing weight, performing in sports and so on. None of these domains had the goal to develop a single grand theory to explain the full range of motivation. It is very hard to encounter a modern theory of motivation these days that does not have a direct practical application or socially relevant use.

In short, the rise of mini theories of motivation is closely connected to the recognition that one grand theory can hardly explain human behavior or is almost impossible to scientifically falsify. The criticism on the grand motivation theories can therefore be summarized in two points. Firstly that motivation theory should not try to explain *all* human behavior. Secondly that motivation theory needs to take cognitive phenomena into account.

2.4.2 Self-determination theory

The result of the cognitive turn and the abandonment of a unified theory of motivation led to replacement of the central role for needs. Most modern theories assume that people perform certain behavior because they believe that this behavior will lead to desired goals. The processes of goal selection and their pursuit is the center point of many motivation mini-theories. Most current motivational research have abandoned the grand theories altogether.

Yet, there is one prominent motivation theory gaining ground in psychology that uses the concept of innate needs as the basis for behavior and uses the grand theories as foundation. Self-determination theory (SDT) maintains that an understanding of human motivation requires a consideration of needs that specify the necessary conditions for psychological growth, integrity and well-being (Deci and Ryan 2000). Goal directed behavior but also psychological development and happiness cannot be achieved without addressing the needs that give rise to these goals and that influences the way the goals are pursued. The focus of SDT is not so much at the *how* of goal pursuit, but at the *why* (process) and *what* (content).

SDT is strongly influenced by the grand theories of motivation that I have discussed. As in the Hull's drive reduction theory, SDT defines needs as "innate, organismic necessities" rather than acquired motives (Deci and Ryan 2000, p.229). In addition it has incorporated the cognitive criticism by classifying needs at the psychological- rather than physiological level. As we have seen, drive theory was very much based on the workings of physiological needs. In fact, Freud stated that the field of biology would someday discover what the real needs of organisms are. SDT holds that these physiological needs exist, but they are given a less prominent place in their explanatory models. The three needs according to SDT are: competence, relatedness and autonomy. It may be clear that behavior according to SDT is mainly psychological driven.

We see that SDT works upon Maslow's humanistic approach to motivation. It assumes that the needs are essential for psychological growth, integrity and well being. These are viewed as necessary conditions for mental health. The satisfaction of these needs is therefore connected with the effective functioning of a human being. This makes satisfaction of needs the key predictor of someones mental health and therefore worth studying - especially in the case of clinical psychology.

Homeostasis and need satisfaction

Although we are not interested in its practical use for therapy, SDT does attempt to explain parts of human behavior with this humanistic framework. In the drive theories that we we have discussed, all behavior can be traced back to disequilibria: the organism strives towards "homeostasis". This makes it vulnerable for cases that cannot easily be reduced to some sort of drive reduction. I have mentioned cases like sitting in a roller coaster or taking drugs at a party, but one can also think of cases that can be attributed to creativity like dancing or solving a crossword puzzle. SDT gives much attention to this criticism - probably influences by its use for therapy.

First it states that actions like these, that fall outside the explanatory scope of the homeostasis, can happen "naturally", without the prod of a need deficit. SDT agrees that much of our behavior is directed towards the satisfaction of needs but once these are more or less satisfied[4], behavior can become more directed towards things or actions that the organism finds "interesting" or "important"(Deci and Ryan 2000, p.230). These things - or objects as Freud would call it - are not always directed towards need satisfaction. Therefore behavior is - once some level of satisfaction has been reached - more or less free to act in the direction of increased psychological differentiation and integration to reach its potential. Thus in terms with its values and desires.

[4]This is also one of the important additions Maslow (1987) made to his motivation theory

Consequently, need satisfaction can be considered a necessity but not all acts are directed towards need satisfaction. Curious exploration of the environment, playing the piano or sunbathing require the nutriments of basic need satisfaction to operate optimally and they can be adaptive as well but these activities are not necessarily consciously intended to satisfy the basic needs.

Second, SDT explains why some of our behavior is contradictory with what you would expect based on the satisfaction of needs. We have seen that drive theory has a hard time explaining cases of self destructive behavior like anorexia or obesity. SDT introduces the concept "need thwarting". An unusually strong desire for food is considered not to be a reflection of a strong innate need, but previous experiences in which the basic need is thwarted. Studies in neuroscience for example, have shown that food disorders can be largely traced back to traumatic experiences in the womb. Food deprivation in the first 7 months is a good prediction for obesity later. Likewise we can explain the strong desire to be in control of the situation. This is not necessarily a stronger need for competence but the consequence of an experience of that need being thwarted. Just as with the obesity example, the observed behavior is overcompensation. Need thwarting is the defense mechanism that changes the behavior into the direction of preventing a previous experience occurring again. It explains why some behavior is the opposite of what you would expect based on optimal need satisfaction.

2.4.3 Evaluation of SDT

SDT provides a well founded solution for at least two problems that drive theories of motivation faced. The first problem is the reduction towards a few physiological needs that cannot explain much of the psychological behavior. Hull was the strongest proponent in reducing all behavior to the physiological domain. Freud already showed more skepticism about this reduction by saying that needs are borderline cases between

the somatic and the psychic. Maslow instead incorporated psychological needs into his model to account for cognitive developments, like competence, prestige, exploration but also intimate relationships. SDT acknowledges that physiological needs are important but states that the focus should be on the psychological domain. Telling is SDT's treatment of typical physiological needs like 'food'. It is not very valuable according to SDT to look at the individual variation of 'food strenght'. It is more fruitful to look at 'self-control' and consumptive patterns: how does the person react to deprivation of food given an environment that is supportive?

Here we also see how SDT tries to meet the second criticism of the cognitive psychologists on the drive theories: that it relies too much on reinforcement and environment. The idea that the function of autonomous men may be taken over one by one as the role of environment and learning is better understood (Skinner 2002, p.58). Instead, SDT holds that some behavior is interesting and important enough in itself. They do not need reinforcement and there exists a psychological class of needs thats explains this.

In short, SDT tries to account for the more complex image of human motivation and the evidence that is provided by the mini-theories of motivation. Therewith it is less ambitious than the grand theories: it doesn't try to explain *all* behavior, but tries to account for certain aspects of it: why do people pursuit their goals? And what are these goals? Why do people behave seemingly in contradiction with pursuing homeostasis? Here SDT seems more directed towards practical use in clinical psychology than the grand theories did. It uses drive theory to explain depression (dissatisfaction of certain needs) or self destructive behavior (needs being thwarted).

Unfortunately, the acknowledgement of the complexity of human motivation and the focus on more clinical terms results in many concepts that are hard to pin down. SDT sees that humans show creativity and competence and therefore assume a "basic, lifelong psychological growth function" (Deci and Ryan 2000, p.232). This

so called growth function never gets seriously operationalized. One of the reasons for behaviorists to leave psychological and cognitive inner states out of their motivational model is that they are hard to scientifically investigate. SDT faces the same problem: How do you measure the need strength for "relatedness"? Or its influence on goal pursuit under situations of stress? As we've seen with Maslow, the seam side of the humanistic approach of SDT is its mediocre scientific method that should support its foundations.

2.5 The universality of needs

One of the great promises of drive theory is the universality of needs. Freud, Hull, Maslow and SDT all argue that needs are innate and therefore culturally independent. Drive theory holds across the globe. From it follows that our endowment is more or less established at birth and will react with the environment around it. This idea however has been disputed, in line with the general debate about the nature / nurture explanations in the sciences. For example, one influential early motivational theorist Henry Murray (1938) argued that needs are learned. Murray approaches needs as a "force in the brain region" (p.123). This is clearly a different conception than Freud, Hull, Maslow or SDT have and for a good reason. Murray's idea was that organisms have 23 needs and their differentiation in strengths explain the differentiation in personality. Social conditioning is the main factor in creating the difference in need strengths and therefore differences between people are largely the consequences of different environments. Defining needs as learned and seeing its strength as a function of learning makes it still possible to conceive needs as universal. However, it gives more much more room to cultural and personal differences.

Although less influential than the view that needs are inherent to human nature, there are many other motivational studies based on idea that socialization uniquely

shapes the causes of difference in personalities and well-being for each person and in each culture especially in the field of achievement motivation (Atkinson 1958) and power motivation (Winter 1973) but also in general motivation studies (McClelland 1985).

To test the assumption that needs are indeed universal, and therewith the universality of the described motivation theories, the psychologists Diener and Tay (2011) asked 60.865 participants from 123 countries about six needs that are similar to those in Maslow's theory of motivation. They concluded that there are needs that are universal to man; that they apply regardless of culture. It follows that needs are likely wired into the human brain and that fulfilling those needs lead to higher well being and therewith counter the idea of Murray's idea of learned needs.

The study also confirms that Maslow and SDT were right on the addition of psychological needs next to the physiological needs. Additionally, Diener and Tay introduce a new need category called social needs. They find that "psychosocial needs" are highly important for "positive feelings" whereas basic need are more important for life evaluations.

However, the so called "hierarchy of needs" of Malsow's model didn't correspond with their findings. When a basic need is unsatisfied, people still get the benefits from higher order needs. For instance, someone who is hungry can still enjoy friendship or be creative. This confirms some of the common examples of some famous persons in history, who with their lifestyle alone question Maslow's hierarchical need model. Rembrandt van Rijn couldn't put food on the table in his most productive years and Jakow Trachtenberg developed his arithmetic system while in concentration camp. Diener and Tay argue that the more basic needs require more attention when they are dissatisfied. There exists a certain order that structures which of your needs are most important but even when that structure determines on what need you should focus first, you can still receive positive feelings from fulfilling less important needs.

Therefore Diener and Tay compare needs with vitamins. They work independently but for well being we need them all.

2.6 Wrap up: What is needed for a computational motivation theory

All things considered and all arguments weighted, if we want to develop a computational theory of motivation it needs to account for the following:

- Needs are innate. They are hardwired in the brain and cultural independent. All people have the same needs and the same need structure. Personal difference are the result of different need values and different satisfaction levels.
- Need satisfaction strategies are learned. They result in habits. Pleasure and displeasure are reinforcers.
- An action can serve multiple needs.
- We probably have physiological needs, psychological needs and cognitive needs.
- There is no pyramid shaped hierarchy of needs. When a need gets dissatisfied it urges for attention. Some needs (like food) may require more attention than others but in the meantime one can still benefit from satisfying other needs.
- Motivation theory has no aim to explain all behavior but since it is an important underlying mechanism of activity and because behavior is one form of human activity it can explain behavior partly and indirectly.
- If we want to proof a motivation theory, we can formulate hypothesis about the expected activity and measure the similarity between the expected and observed activity (Hull).

– There is no definite list of needs and motivation theory should not aim to make one. There are however overlapping needs that seem to be universal (Freud, Maslow, Diener & Tay).

Chapter 3

A computational version of motivation theory

3.1 MicroPsi

One computational attempt for motivation theory is currently in the making by Harvard researcher Joscha Bach. Bach is the main developer of the cognitive architecture MicroPsi (Bach 2012, 2007). Cognitive architectures attempt to give a blueprint of the mind. They are, to quote one of its pioneers "a specification of the structure of the brain at the level of abstraction that explains how it achieves the function of the mind" (Anderson 2009). In this sense they are still on the original quest of AI: understanding how the mind works by building a working model of it.

Well established cognitve architectures are Act-R and SOAR with each their own theories and assumptions about how the mind works (Kotseruba et al. 2016). What distinguishes MicroPsi from other existing cognitive architectures is that it tries to answer the question how cognitive processes and behaviors give rise to cognitive autonomy, personhood and phenomenal experience. Other architectures often restrict themselves to how human behavior can be simulated within a technical framework.

MicroPsi wants to know how a mind gets into being; it is a framework for thinking about the mind. It tries to develop tools to answer questions about the philosophy, functionality and the physiology of cognition (Bach 2007, p.227).

Agents are the embodiment of the architecture. MicroPsi-agents are primarily general learning systems. This perspective suggests that behavior and perception are emergent and largely based on probabilistic models. Intelligence is seen as making models of the self and of the world and different parts of the brain allow for supplying a complex and hierarchal infrastructure for cognitive processes like attention control, learning and pain and pleasure modulation. It follows that we need to understand how this emergent behavior comes to be; how the structures of cognition like problem solving, imagination, perception, reasoning, motor movement, memory, social interaction, language and so on are shaped. The underlying foundation that gives MicroPsi-agents its direction and energy must be found in its motivational responses to the environment.

Like most other cognitive architectures, MicroPsi is still much work in progress. Right now, most headway has been made in simulating a spreading activation network where it uses neurosymbolic and monolithic representations in the form of node nets. The node nets form a layer that acts between perception and action. Everything that controls MicroPsi- agents has the same structure and by using these representations we can learn something about planning and learning. However, not much progress has been made in the question of what makes agents learn and plan in the first place.

Since MicroPsi assumes that behavior and perception is emergent, it has to define and explain how general learning gives rise to cognition and intelligent agents. This is proposed as being a motivational system. This system, the thing that makes MicroPsi so unique and interesting, has yet to be developed and implemented (Bach 2017). So although still in embryonic phase, here I present how such a motivational

system would look like according to Bach[1] and in what way MicroPsi's computational motivation theory matches the requirements of a working motivation theory as laid out in the former chapter. The to be implemented motivational system is what I call "the motivation machine".

3.2 The motivation machine

The motivation machine starts with a predefined set of demands of the system that are called *needs*. These needs are the basic elements of motivation. All behaviors are either directed on the satisfaction of a need, or on the avoidance of the frustration of a need. What the agents has to do is avoid, pursuit and consume to survive and thrive in its environment.

3.2.1 Three needs

Three categories of needs are distinguished: physiological needs, social needs and cognitive needs. The **physiological needs** are the somatic needs that regulate the basic survival of the organism; if an organism would be alone on the planet, he would still have physiological needs. They are defined as "food", "water", "health", "rest" and "libido". They give rise to foraging, eating, drinking, sleeping, walking, mating and so on.

Social needs

The social needs direct the organism towards interaction with other individuals and groups.

[1]Bach's ideas on motivation has changed in time and are often raw sketches without much foundation in literature. The computational motivation theory that I describe is distilled from much of his works - some published and many unpublished papers, talks, code and especially from the many hours of conversation that we've had.

Affiliation is the need for recognition and acceptance. The need to belong. It is virtual currency for collaboration: it gives rewards for cooperation and punishment for defection. It involves social signaling like frowning, smiling, raising fingers which enforces cooperative behavior.

Legitimacy is what makes the organism act conform to its internal norms in the absent of direct observers. It is the honor of the organism.

Nurturing makes the organism care for others. The need gets satisfied by subjectively increasing the well-being of other individuals and groups. Recipients that are closer give higher rewards than more anonymous recipients.

The need for *Dominance* makes the organism want to climb in social hierarchies and care for the established position. It defines ones competitiveness and need for power.

Affection makes one want to bond with some other individual in a romantic way. This can lead to closeness and courtship. It does influence - but is different from the physiological libido that is directed to sexual behavior and can be satisfied alone or with multiple partners. Affection makes one want to bond with one person in particular, hence the saying: "he/she is the one".

Cognitive needs

The cognitive needs are related to skills, play, creativity and discovery.

Exploration makes the organism want to acquire certainty about the environment. When it encounters and recognizes objects or processes, certainty gets increased and the exploration need satisfied. So for example, when a organism gets into a new environment, one of the first things it is likely going to do is explore it surroundings.

Competence is a special need. It indicates how much the organism feels able and suitable to solve a problem: "how effective am I"? Competence gets satisfied by efficiency signals from other needs. When they are easily satisfied the organism feels

competent. Competence is regulates risk taking. So the idea of a low competency in a certain field has a negative influence on risk taking.

Aesthetics is the need that drives the organism to look for patterns within the knowledge it accumulates, to represent or receive it more compactly. It often includes stimuli that are in and of themselves pleasant like harmonious sounds and it involves the formation of mental representations and the drive to identify the structures in them. Aesthetics is a special need too because it distracts from the immediate needs of survival. It is responsible for mathematic elegance or for the the poet consumed with finding *le mot juste*.

3.2.2 How do needs work?

Remember that the organism doesn't have pre-defined goals. The reason for drinking a glass of water in the morning is not because there is pre-defined goal to drink water every morning but because it satisfies the pre-defined physiological 'water' need. This means that agents establish their own goals based on its corresponding needs. How does that work?

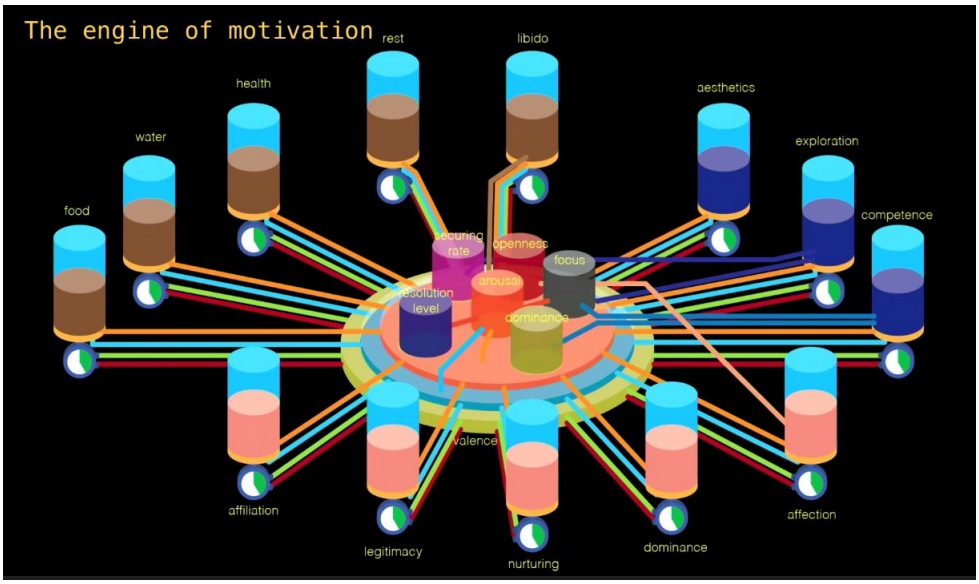
Each need is normalized between 0 and 1, and its corresponding *urge* and *reward* signals are weighted by a *strength parameter*. A need functions like a tank. When you sweat, your water tank empties and the value of the water need lowers. Needs have a natural *decay factor,* meaning that they deplete over time independently; even on your most lazy days you will get sleepy, hungry and thirsty that indicates that the rest, food and water demands increase.

The *weight* of a need defines how strong the need is compared to other needs. This is important because the need structure of needs is not hierarchical. Their formation is not pyramid shaped in which some (basic) needs get satisfied first before other demands start drawing for attention. All needs are mutually active but some needs' weights are higher, causing them to signal for more care than other needs. When

a need becomes dominant, that is, when the need value crosses a certain threshold, the need gets signaled by an *urge* indicator. The higher the urge, the stronger its demand for attention. The urge is the difference between the need's target value (a fully satisfied need) and its current value. On top of the urge, a need is also characterized by *urgency,* defined by the necessity to satisfy the urge. This necessity is determined by the time left to get satisfaction. Sometimes a weak need might have a short time window to get satisfied which makes it more efficient to attend first. The same occurs when the needs depletion is suddenly high, for example when a crucial resource is about to drop.

Satisfaction happens as a consequence of a *consumption* event like drinking a glass of water. The need value - in this case 'water' will increase and the urge will decrease. *Gain* and *loss* determine how easily the need gets satisfied or frustrated. When the gain value is high, the effect of the consumption event is high too. The same holds for loss.

Pleasure and pain are the main learning mechanisms of the organism. They give rewards and make reinforcement learning possible. Pleasure and pain are connected to the increasing and decreasing of the need values. Satisfying a need leads to pleasure signals and frustrating a need to pain signals. An empty stomach is a depleted need for example that can lead to a pain signal. The length of the pleasure and pain signals has a natural decay, until the maximal signal disappears. In addition, depleted needs may also cause enduring pain where pain signals are sent as long as the need value is within a certain range, for instance between 0 - 0.1.



3.2.3 Consumptions and events

The needs drive the organism towards certain behavior that is aimed at reaching homeostasis. The organism tries to keep its needs balanced using feedback loops that constantly track the difference between the target and current value of the need. To satisfy a need, the organism has to "consume" something. This can be objects or actions like eating food to satisfy hunger, sleeping to get rest or walking and looking around to explore the environment. So a consumption is the part of an event that influences the need(s) of an organism. When an event gets triggered, the consumption generates an amount of satisfaction or frustration for an associated need. For instance, the eat consumption generates satisfaction for the food need for a certain amount of time. This time window is limited by the maximal reward of the consumption event: eating an apple will create less satisfaction for the food need than an whole meal and eating a whole meal is different from eating 10 apples. The strength of a consumption is regulated by its signal strength that plays the function of a neurotransmitter. The idea is that the signal starts very strong and decreases slowly reflecting an immediate reward function. A consumption also has a duration.

The motivation machine has different consumptions which are associated to a need. For the *physiological needs* they are eat, drink, sweat, recover, sprint, bruise and mate. For the *social needs* they are acceptance, rejection, pride, shame, support, supplication, win, loss, connection and abandonment. For the *cognitive needs* they are success, failure, confirmation, disconfirmation, admiration and disgust.

Events

It is said that when Harold Macmillan became Britain's prime minister in 1956, he was asked what would determine his governments course. He replied with Edwardian languor: "Events, dear boy, events." In the motivation machine events come in two categories: they can just happen or they can be formulated goals of the organism and actively pursued. Important to remember here is that the events are part of the motivation machine of the organism and not of the outside world. Many events do not appear out of thin air but they are anticipated situations. These *expected events* hold into account that most situations that happen in the perceptual world of the organism are similar to earlier situations or are foreseeable. Furthermore, many events happen in the world but only a few are relevant for an individual organism. Someone turning the light switch 10.000 km away has no direct motivational relevance to you. Events are only relevant when they are associated with the expectation of a consumption (like exploring), i.e. with satisfying or frustrating a need; you will perceive an apple differently if you want one. This means that the values of an event don't correspond with the sensitivity of the agent but to the extend to which consumptions activate an event. In addition, events can be appetitive or aversive. In the former case they give a positive reward and the goal is to pursue the event and in the latter a negative reward with the goal to avert it. The *certainty* of the event specifies the confidence that the event will happen.

The expected events are represented in a list. At the end of every loop (more about this later), the motivation machine triggers *belief updates* which establish, change, execute or delete the expected events in the list. When an expected event changes due to one of the belief updates or because the agent visits the event, it generates *expected rewards* depending on the *certainty.* These rewards work as expectation managers and help the agent with pursuing or avoiding the event. If an expected event manifests, the agent's certainty increases that is associated with the satisfaction of the exploration need. If it doesn't manifest, it frustrates the certainty and thus increases the demand for the agent to explore its environment. If the events are *goals,* the manifestation of an expected event or the absence thereof affect the competence need of the organism.

3.2.4 Goals and motives

In the motivation machine needs determine what has to be done. In this schema, reaching a goal is satisfying a need. Because the organism's goals are not pre-defined, they have to come from the bottom up. In this sense, what we are looking for is a *motive* for the agent. A motive can be defined as a goal, an urge and a sequence of actions to reach that goal together. How does that work?

Identifying and pursuing goals and corresponding actions are the result of a decision making process. The first step in the decision tree is an activated need. A need gets activated through a consumption or due to natural decay. Subsequently a need value can cross a certain threshold that triggers an urge signal. This turns the need into a demand. Through past learning[2] the urge is associated with actions that has satisfied or frustrated the need in the past called an *appetitive goal* and an action that has frustrated the need is called an *aversive goal.* We are now on two third of

[2]a protocol memory has yet to be implemented. This is already part of MicroPsi (Bach 2007) but not part of the motivation machine

reaching our motive: we have a goal and an urge together. The next step is finding the corresponding action to reach the goal.

The first thing the agent will try to do in order to satisfy the active need is execute a plan that has worked before (a habit) and turn them into actions by executing the plan. Most of human behavior is automatized. If the agent is hungry and it wants to get an apple from its backpack, the procedure to open the bag to get to the apple, grab it and close the bag again is an automatism. Habits can be found in the agent's memory and are the result of learning. If such a strategy can not automatically be found and executed the agent will try to allocate resources to develop a new plan. By matching the existing plans and world situations and actions into a new one it tries to connect the current situations with the appetitive goal. If plan construction fails, so when none of the existing material matches the demands from the current situation, the agent gives up pursuing the current urge. When a motive is in the situation with no plan found to satisfy, the motive can become frustrated. The agent needs new information to come up with a new plan. Therefore, a frustrated motive increases the exploration need and depending on the frustrated *motive strength* (a combination of an expected reward, the urgency of the resolving need and the competency of the agent), the agent will start exploring the environment for new information to deal with the situation at hand.

A plan that has been correctly identified amounts to a motive. If a motive becomes active it is not always selected immediately. Only one motive can become the *leading motive:* the motive that is the strongest. This is the most dominant motive that requires the most attention of the agent and that governs the actions of the agent. A leading motive can be dropped at any time nonetheless and be replaced for another active motive. Imagine an agent with a leading motive for resting. At the same time it has an active (but not leading) motive to hydrate. This means that the need for water has crossed its threshold whereupon an urge signal has been triggered.

The need has become active and the agent will be looking for a plan to satisfy the water need. However, apparently the motive for resting has gone through the same process and somewhere in the computation the values of the rest need are in higher demand for resolving: the rest motive is stronger than the hydrate motive. The agent is executing a plan he has found in its memory to get to bed and developed a corresponding strategy. However, along its way to bed, the agent crosses a water fountain. The agent recognizes this event as a drink consumption and the agent's motivation machine calculates if the expected reward in combination with the strength of the motive if it is worthwhile to visit the event. If it is, suddenly the weaker active motive to hydrate becomes leading because it exceeded the selection threshold that determines how easy it is to switch between motives. The leading motive to rest will be dropped an become active again while the motive to hydrate is leading now until satisfied.

3.3 The motivation machine compared

The motivation machine is a computational model of motivation theory. One of the advantages of computational models compared to purely theoretical models is that they have to be precise in order to work. The motivation machine is already precise and before I describe how to make him work, let me review the choices that have been made in the current motivation machine. How do they compare to the aforementioned motivation theories? One of the disadvantages of a computational model - especially in this case where a brand new model has been build from scratch- is that to make it scientifically relevant, much theoretical backing has to be given for all the many different individual choices that have been made. What needs do we define? What is the value, weight, gain and loss value of every individual need? what is the consumption value of every consumption. Here I would like to to both:

reviewing the choices and see if they have theoretical backing and compare these choices with the discussed motivation theories.

3.3.1 Categories of needs

In the motivation machine, the behavior, action regulation and decision making of the agents are founded in an architecture of physiological, social and cognitive needs. These needs are the foundational framework of the motivational model. That motivation is based on needs is as we've seen the standard in most classical motivational theories. Freud's drive theory, Hull's drive reduction theory, Malsow's hierarchy of needs but also the modern motivation theory of Deci & Ryan have a list of needs as the underlying mechanism of the motivational system.

The categories of needs differ among the different theories and also between them and the motivation machine. Freud and Hull claimed that all motivation could be derived from a few physiological needs and believed that the fresh field of biology would discover them in the future. It should be said that Freud shows much hesitance about clearly defining the needs since he admits that he has not enough knowledge about them. He defines them as a borderline case between the cognitive (psychic) and the physiological (somatic). We have to conclude that the field of biology hasn't come up with a definite theory of physiological needs.

Maslow stressed explicitly that the physiological needs are not the center point of motivation. He incorporated social needs in his early theory (1943) like "intimate relationship" and "friendship" on top (literally) of the physiological basic needs. Confusingly, he placed them in a broader category called psychological needs. Later, in his revised theory of motivation (1987) Maslow added a third categories of needs: cognitive needs.

The definitions of needs and the categorizations in Freud, Hull, Maslow and modern theories has a sway of arbitrariness. This is also the case in modern theories.

Deci & Ryan and SDT characterize three needs: Competence, relatedness and autonomy and add that we probably also have physiological needs that they further ignore, probably because of SDT's clinical orientation. Diener & Tay define six needs. Our motivation machine has 13 needs.

What has stood the test of time is that we seem to have both physiological and psychological needs. The motivation machine goes one step further, along with Maslow and Diener & Tay and add a layer of cognitive needs (competence, exploration, aesthetics). How the needs are defined precisely will be important in the future, but right now it seems that none of the motivation researchers has a clear definite list of needs. To show that intelligent organisms are indeed motivational and to develop such a theory of motivation, it is right now not necessary to know exactly what needs we have. This will develop in future implementations and experimentations with the architecture. Nonetheless, the motivation machine has two needs that are important for the system, but stay unclear and deserve discussion.

Competence

The need for *competence* is a questionable one. Competence is the need that measures how able the agent is at solving the problem at hand. When choosing an action, the agent weights the strength of the corresponding urge against the change of success (Bach 2007, p.50). Herewith, competence is the main risk calculator, the competence amounts to something like self-confidence of the agent. Competence is also one of the most important concepts in Psi theory, the main theory that is the main inspiration for MicroPsi (Dörner 1999, Dörner and Güss 2013). The competency tank is filled with efficiency signals: the more successful the agent is in satisfying it's needs, the more competent it feels. This means that the competence tank is connected with all the other tanks. If the food need fills, the competence need fills as well; if the agent feels satisfied, it feels competent.

We can already see the difficulty here: the competence need functions at two different levels. A kind of Andre Rieu: a conductor of an orchestra who also plays the first violin. The conductor is satisfied when the orchestra plays well but the playing well of the orchestra depends on the performance of the first violin (also the conductor) and of the performance of the conductor (also the first violin).

Having a competence need is defensible though: this means that the agent wants to be good at what it is doing. The typical nerd for instance has probably a higher than average demand for competency and lower than average demand for affiliation and dominance. The point here is that this should be unconnected to the satisfaction of the other needs: the nerd is not more competent or satisfied in general - what could perhaps be called general happiness instead. The nerd does worry about his/her social skills and there is no reason to think that the nerd doesn't want to belong or has more self-confidence. Therefore, having a competence need functioning just like the others would make more sense than having it functioning at two different levels as has currently been proposed in the motivation machine.

Aesthetics

The other interesting need is *aesthetics* (looking for patterns). In none of the other theories we find a resemblance of this need. It functions as the need that drives the agent to do things that are worthwhile in an on itself, like making music or enjoy the elegance of a mathematical formula. Bach doesn't support why we have this need but I think that the need for aesthetics is indeed defensible.

First of all we have to be careful with the placement of aesthetics. Since were are covering the motivation machine, aesthetics has been placed as part of the cognition of the organism. Just as with most of the subjects of philosophy, the thinking about aesthetics has been subject to a continuously dialectic between the objective and the subjective. To prevent myself from dwelling into all the epistemological assumptions

that must have preceded the choice to put aesthetics *inside* the mind, let us ignore the difference between aesthetics in the world and in the mind for now and let us see what it means for an agent to have this need and if it can be justified.

Placing a need for aesthetics in the mind means that aesthetics is (at least) subjective. The need for beauty is in the eye of the beholder. It covers the creativity and spontaneity of the individual. As opposed to science or history, aesthetics gives the agent the opportunity to enjoy something purely particular. Aristotle said that drama gives us one action that is a unity in itself. Science and history on the other hand always cover the universal, their goal is to generalize over many things instead of staying with one particular thing. The German philosopher Ernst Cassirer illustrated this in his *Essay on man* by letting two painters paint the same landscape (Cassirer 1944). What we get are two very different paintings. After all, an artist doesn't try give us an exact copy of the empirical object. He does the opposite: art and in the expressive in general as the French philosopher Jean Jacques Derrida claims is "pregnant" of infinite possibilities that remain unrealized in the common experience of the senses. An artist tries to capture a conscious state; a feeling of what it is like.

This gives answers to questions why we build "unnecessary" complex, tedious and expensive architectures, have a strong preference for symmetry or get goose bums while hearing a perfect harmony, without any social economic reward or reward for survival. In an interview this year Bach[3] speaks briefly about the need for formatting mental representations when he speaks about qualia. He thinks that our mind needs to look for abstract mathematical structure to learn language. This mathematical structure that is needed for grammar is something that we can not directly observe. To look for this structure and to learn it we need a drive built into our mind that receives reward signals when we find interesting compositional structure, like we do in experiencing and enjoying aesthetics.

[3]https://www.youtube.com/watch?v=noScrzeEh4o

Now, one could argue, the scientist also tries to look for structure. How is that different? The scientist is looking for truth and tries to make a model of the world, preferably one model that explains everything. The artists however doesn't care about truth. He doesn't need to make an accurate model. Like the painter of Cassirer, he needs to capture a particular conscious state and therefore needs to change his own brain state into one that is in a state of phenomenologically experiencing it. Therewith, with creating something new, the artist can share this experience with someone else.

Aesthetics in this sense is useful because it helps us making sense of the world around us. Herewith aesthetics is closely related with how Immanuel Kant describes it in his *Kritik der Urteilskraft* (Kant 1963). Kant proposes that aesthetics has a function in understanding the world around us. It is one of the conditions of our senses to perceive to world. This perception for Kant is not passive: with a sudden change in his frame of mind, the artists can see the same thing with different eyes. It is not occupied with the immediate true things, but with the unknown possible forms. "in such absorption in the dynamic aspect of form consists the aesthetic experience"(Cassirer 1944, p.194).

This also answers to the concern of SDT. SDT tries to account for active personhood that - as they claim - can not be established with a deficit based motivation theory[4]. How can that account for curiosity? Or for a child that enjoys playing the piano, and why we can feel ecstatic when we discover something perfect. With the need for aesthetics Bach tries to solve this issue. Satisfying the need is perhaps helpful as a byproduct for something else[5], but the core of it is that it is not directly beneficial for our survival. It can be enjoyed in and on itself. As a particular sensation.

[4]see 3.3.3

[5]In a skypecall Bach said that he wondered if aethetics is a byproduct of language acquisition.

3.3.2 Homeostasis

An important concept in classical motivation theory is homeostasis: the tendency of the organism to keep its inner environment stable[6]. An important implication of homeostasis is that most of the agent's behavior is driven by this inner regulation system that is continuously searching for an equilibrium of all the need values. The following implication is that a fully satisfied agent - one that is in the equilibrium state - falls into inertia: it won't get out of bed in the morning.

Of course the equilibrium state is hypothetical for the reason that needs have a decay factor alone. So the supporter of this idea (like Clark Hull and Sigmund Freud, Dietrich Dorner and Joscha Bach) can therewith explain that we usually don't encounter people without any activity and thus without any motivation. Perhaps some animals are in this state, especially solitary animals in an environment that doesn't give them stress. I have two cats: Henk and Iggy. If the definition of stoics is that you only worry about those things that you can change, Henk and Iggy are stoics. They are fine with themselves and they are fine where they are. They only get get moving when they really need to: when they get hungry, when they want to get pet or when they have to go outside for a little bit of stretching and relieving. Perhaps Henk and Iggy are most of the time in their equilibrium state.

Human beings however are not. We are not fine in the morning because we have all these needs to attend. Even a Buddhist Zen-masters verifies the homeostasis idea. Not because the Zen-master knows how to keep his needs permanently satisfied. That is impossible. The Zen-master has found a way to hack the motivation machine's reward function. When a need gets frustrated, it sends displeasure signals: "don't do more of this. Do something else instead". Only when you are able to fool your reward system, you are able to reach a sense of enlightenment[7]. This adjusts your

[6] see 2.2 and 2.4.2

[7] in German philosophy you have to concepts for enlightment: *Aufklärung* and *Erleuchtung*. Aufklärung is enlightment in the famous *Sapere Aude!* sense: the idea that you should think for

motivation: it fixes what is relevant for you and how you relate to this and it changes the relationship between self and universe.

The important implication that is neither mentioned by the classical motivation theorists, nor by Bach is the passive image of the organisms. That shouldn't be a surprise: the purpose of drive theory was to explain how organisms went from inactive to active. Without motivation we would all be lethargic. The concept "motive" itself comes from the Latin *movere*, which means: "to move". Motivation theory in this sense is the study of energizing the passive and the whole purpose of the motivation machine is to direct and energize behavior.

This passive image however is not self-evident. Just as the natural state of planets is in motion, it could be argued that people are too. Maslow (1987, 2009) for example wrote:

> Sound motivational theory should [...] assume that motivation is constant, never ending, fluctuating, and complex and that it is an almost universal characteristic of practicality every organismic state of affairs

And Deci & Ryan(Deci and Ryan 2000, p.230) wrote:

> In SDT, the set point is growth-oriented activity. That is, rather than viewing people as passively waiting for a disequilibrium, we view them as naturally inclined to act on their inner and outer environments, engage activities that interest them and move toward personal en interpersonal coherence. Thus, they do not have to be pushed or prodded to act. Further and importantly, their behavior does not have to be aimed at need satisfaction per se, it may simply be focused on an interesting activity or an important goal if they are in a context that allows need satisfaction.

yourself. Aufkläring is very much connected with rationalism. Erleuchting however is used for the Buddhist *nirvana.* What I mean here is *Erleuchtung:* The Zen-master hasn't fixed his rationalism, he fixed his motivation system.

The difference between the active or and passive image of the organism is subtle but important. Instead of dealing with deficit motivational concepts like tension, reduction, homeostasis and equilibrium, the active perspective deals with growth motivational concepts like creativity, competence, curiosity and self actualization. The latter emphasizes the cognitive constructs that appeared as a result of the cognitive revolution in motivation theory[8] like plans, goals, expectations, belief and the self-concepts, while the former emphasizes the physiological and environmental constructs. This active image reflects the influence of the humanistic tradition in psychology as we have seen in Maslow (1987, 1943). Not surprisingly humanistic psychologists resist the machine metaphor that portrays motivation in a deterministic fashion.

Interestingly enough, the motivation machine is a combination of the two positions: it sees the nature of the agent as passive but it emphasizes the cognitive constructs. There are two explanations for this position. Firstly, the deficit motivational concepts are better formalizable in a computational model than growth motivational concepts. How do you make a model of a person that "may be simply focused on an interesting activity" as Deci & Ryan describe it or one where children are "unendingly curious"? (Reeve 2009, p.37). Perhaps that these concepts are useful and clear enough for the purposes of clinical psychology and psychotherapy but for the purposes of AI they are too opaque to model, let alone to program.

The deficit motivational concepts that give a passive human image are more systematic. For instance, *tension* in the motivation machine can be translated with the combination of urge and urgency that determine the strength of the demand for attention. The higher the urge and the urgency, the more resources will be deployed to relieve the tension by satisfying the needs. *Equilibrium* in turn simply means no

[8]See chapter 2

cumulative difference between the target value and the current value of the needs and so on.

The danger here is "methodologism": the fallacy of ignorance towards those aspects that happen to fall outside the methodology at hand (Feyerabend 1975). It is not the methodology that should determine what theory one develops; instead, it is the question one asks that should find the methods for answering it. Giving a passive human image because it fits the methodology of computation can not be a legitimate response to the active image. Since no other reason or defense has been given yet by Bach or others, the only way to test the theory is by testing the model and see if needs like competence, exploration and aesthetics do lead to the desired outcomes that Maslow and SDT try to account for.

3.3.3 The specific configurations of the motivation machine

When it comes done to the specifics, the detailed configuration of the motivation machine has largely been set, based on common sense. This means that for example the values for the needs, like the initial value, the weight, decay, gain and loss are picked on what Bach thinks that makes sense. For example, the weight of the food need is 0.6 and that of the water need 1. Therefore the agent can live longer without food than without water. The mutual ratio however is up to debate. This applies to the specific values of consumptions and events too. How long it takes before an agent consumes and apple in an eating event, and what specific consequences this has for the urgency of the corresponding needs should follow from comparison with actual research and by implementing an testing the motivational model in an actual environment.

Chapter 4

Implementing the motivation machine in a virtual world

4.1 Why we need a virtual world

4.1.1 A motivated agent in a controllable environment

One of the criticisms on Freud's theory of drives and Maslow's hierarchy of needs was their limitary research methods. Both based their generalizing grand theories on non representative samples and uncontrolled environments. Hull and other behaviorists tried to solve this by designing laboratory experiments with rats. The main advantage of these experiments was that the researcher could control for external stimuli that may influence the internal state of the organism. However, the shortcoming of the behavioristic approach was that they ignored - and in most cases even denied internal states altogether. Here we saw another instance of methodologism: internal states fell outside the behaviorist' scientific toolbox and got disregarded. On the other hand, the methodological approach itself to perform controlled laboratory experiments has stood the test of time and has become one of the main scientific methods today.

One of the main advantages of a computational theory of motivation is that we have access to the observations of the internal states in the form of data. Therewith we don't have the problem that Freud and Hull faced of inaccessible and unknowable internal states. So what we want is a controllable environment in which we can test an organism with a version of the motivation machine: a motivated agent. We want to know and see how this agent interacts with the world based on his motivational system. We want to know if drive is predictable by manipulating the environmental conditions and changing the internal state of the organism - as Hull tried to do. We want to make expectations based on the configurations of the motivation machines in combination with the stimuli of the outside world and see if the agent behaves in correspondence to our expectations.

4.1.2 External stimulation

If we we want to say something with our experimental design about real organisms in the real world we need a model of internal and external stimulation.[1] The motivation machine deals with an essential part of the internal stimulation. It is an effort to build a real motivational system of a human being on a functional level that simulates as specific part of the internal processes of that human being.[2]

External stimulation is the other side of the same coin and at least as complex. We can start simple by saying that we exist in a world and that we are always in interaction with it. We can extend the conventional assumption that what we get when we are in the universe is information (Dennett 2018). But one way or the other we enter the very troubled waters of knowledge theory that offers way too much thought than can be dealt with here from *schematism* (Kant 1987) to *sense data*

[1] I am not going to debate the topic of strong and weak AI here nor the topic of *multiple realizability* that covers if the core of human thinking is symbol manipulation, what is " real" stimulation and so fort. I go along with the standard assumption in AI that the world consists of information and that our mind is an information processing unit, a symbol manipulator.

[2] Of course there are more internal processes going on, but although many are connected to motivation, they are not the core of this thesis.

(Russell 1912) , from *logical empiricism (Quine 1951)* to *conceptualism* (McDowell 1996) and way beyond. Unfortunately this would stop us from making any progress right now so let me present a very rough outline of my idea of this external stimulation without going into too much philosophical detail and with relevance for implementing the motivation machine.[3]

Information theory

The world is a computational chaos and consists of information. The existential characteristics of this information remain unclear to us, but we do have a method to split up the information that our mind receives via the senses into uniform quantities, called bits.[4] This method was introduced to us by Claude Shannon (1964) and works for any information processing system. It refers to the statistical relation between different word states: what information about state A can we derive from the observation of state B? State A must be in some form in a causal relation with state B and this causality can show variation. It follows that the meaning of information is change. Our mind is a system that makes models by identifying relationships between changes in information. What we get from the world is not objects, energy, matter or light but discernible differences and the name for that is information.

Our mind receives this information in the form of patterns via the senses. We can resolve these differences into little yes' and no's and maybe's and create percepts from it. We probably use the same low level input as machine learning algorithms do but instead of filtering out the invariance until we have something left that never changes; we model a complete world. Therewith our mind is not a classifier but a simulator. Our picture of the world seems to be a simulation and it is correct that

[3]Thanks to Joscha Bach in correspondence

[4]These bits don't need to be digitalized; long before we had computers, we were able to save, signal and process information.

organisms with different brain states and different perceptive capabilities perceive the world in different ways.

Grounded systems

With this idea of external stimulation we are equipped to counter objections that have been categorized under the umbrella of *the symbol grounded problem.* In the 1980's the philosopher John Searle (1980) started a lively discussion surrounding his famous Chinese Room argument: how do symbols used in an AI system refer to meaning? The idea is that using a simple finite set of symbols isn't enough to capture the richness and heterogeneity of the real world. Computational systems can not generate their own semantics while humans can. This has lead to a movement in AI where researchers expressed the idea that information processing systems without contact with " the real world" would never be intelligent (Brooks 1991). As a consequence this *nouvelle-AI* movement said that to require meaning we need physical bodies in a physical world: embodied AI or robotics. The only way to represent and manipulate grounded symbols, they claim, is by letting agents interacts with the real world using body parts instead of only doing symbol manipulation in the mind. This leaves only robotics up for the task to reach AGI.

There is some truth to this idea: since we are aiming for an autonomous agent, our AI-system must be capable to autonomous make sense of the world, just as we humans do. Cognition may be a product of the mind but it is clearly situated in a world.

Nonetheless, I don't see why this world should be a real world. What matters is not that the world shares the same atomic make-up but that the system is able to derive similar representations on a functional level. Since we are satisfied with models that work on a functional level, we are looking for a world that simulates our world in way that is congruent with the way information is processed by the brain

on that functional level. If we could hard code the exact same representations as in the real world, this would not somehow be "semantic inferior" to those acquired autonomously.(Bach 2007, p.178)

Either way, the agent has to be able to receive external information and transform it into a world model from which it can make plans, actions and find goals to satisfy its needs, just as we do. Of course, the more the information of our simulated world is similar to our world, the easier we can generalize conclusions from our implementation. We should therefore aim for grounded systems that do interact with a grounded environment but not get entangled in the symbol grounded problem.

Shortly, there are strong reasons to implement the motivation machine in a virtual environment. We want to say something about the mind that is situated in a world. That requires from the environment that it strives to simulate the richness and heterogeneity of the real world and from the mind that its representations refer to the structures of that environment. We need a suitable set of patterns to constrain the representations similar to ours. Furthermore, we want an environment that is controllable and has the benefits that Hull aimed for in his experiments without the long and costly road towards robotics. We want to be able to manipulate the external stimuli to control and measure its effects on the agent to see what part is caused by the variables of the motivation machine and how. Therewith we can perform experiments that simulates experiments in the real world and that has access to the internal states of the agent and draw conclusions from it.

4.2 Minecraft

One virtual environment that is suitable for our task is the computer game Minecraft. Minecraft is based on the simple concept of a world entirely consisting of standard sized building blocks that is 3D procedurally generated. In this world the player of

Minecraft gains knowledge about its environment and can built about whatever he or she likes from textured cubes which are arranged in a fixed grid pattern and represent all kinds of resources and materials.

Because Minecraft's world is build algorithmically instead of manually it makes it more realistic than many other video games. Objects share a similar general structure but its particular characteristics are generated randomly which gives every tree, mountain or other object a unique character. Therefore, just as in the real world, a Minecraft world has many different biomes like plains, tundra, desserts and forests and many other characteristics that are neither hard-coded nor totally randomized.

Minecraft is also suitable for implementing our motivation machine for practical reasons. It is easy to play, cheap and it has a lively community with comprehensive documentation. Furthermore, it has many options to set up challenges and tasks that are similar to real world tasks like navigation and planning. It also supports interaction with multi agents and human-played agents interacting with AI-agents.

Minecraft also has the potentiality to express the drive structure of the motivation machine. A Minecraft agent has basic physiological characteristics like health and hunger. The world provides corresponding ingredients like meat, apples, milk, honey and so on that can be consumed. This requires moving, collecting, building, learning and planning. Altogether this gives great opportunities for AI to interact with the environment. That can be on the low-level of cognitive processes like seeing, hearing, moving around and on higher-level processes that involve intelligence. Therefore Minecraft suits as a good virtual environment for an AI-agent, or in our case, a motivated agent to do experiments with.

4.3 The implementation

4.3.1 Malmo

Implementing the motivation machine in Minecraft works by using Malmo. Malmo is an open source platform for AI experimentation developed by Microsoft (Johnson et al. 2016). It provides an abstraction layer on top of Minecraft that instruments to expose an API by which agents are integrated to an environment through a sensorimotor loop. The information from the Minecraft world can therewith be received by the agent and the experimenter can construct increasingly complex tasks and perform experiments.

Malmo sets up a Minecraft host server which gives control over Minecraft and in which the researcher can run programs. The barest skeleton of a mission provides the user with to following concepts:

The *MissionSpec* sets up the mission for the researcher. Any environmental world can be set up by providing the MissionSpec with an XML-file with specifications. It can equip the agent with tools and items, determines what objects are located in the environment and can attach rewards in the case of learning.

The *AgentHost* mediates between the researcher's code (the agent) and the world. It instantiates the mission and sets up an agent whereupon a mission gets started. During that mission the researcher can send commands through the AgentHost such as 'jump', 'move', 'crouch', 'attack', 'use' etc. It can also request as *WorldState* through which the agent receives information about the world and the self in that world. This can be both sensory as direct information - depending on the video policy. Information about the self is very limited however, like 'time alive' and 'coordinates'. That is resolved by the motivation machine.

4.3.2 Implementing the motivation machine in a Minecraft Agent: creating Adam

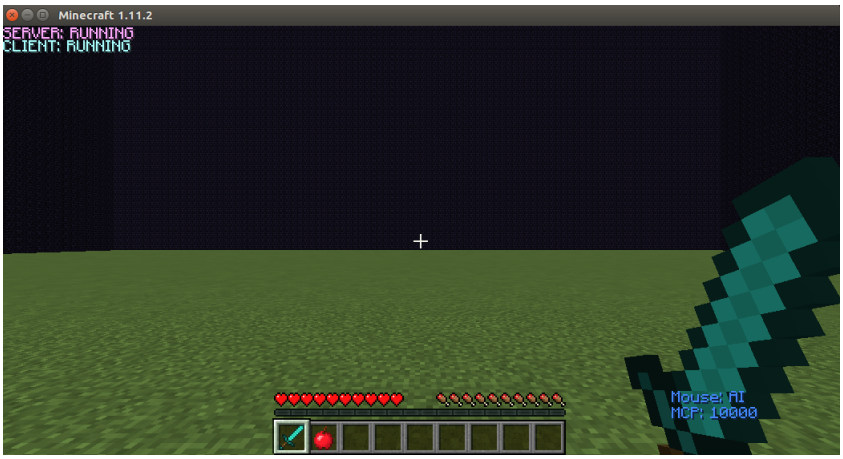
To implement the motivation machine I initiated a class that creates default Malmo objects that correspond with the agent's needs and consumptions. When a mission starts, these needs and consumption configurations get parsed into the AgentHost and they form the motivational state of the agent.

Using the *FlatWorldGenerator,* the *MissionSpec* loads an XML that procedurally generates a 1 hectare field of grass that is surrounded by a wall using the *DrawCuboid* for not letting the agent walk out of the experimental ground. A Python script that codes an object constructor inserts apples and bread randomly on the ground. The agent - called Adam - gets spawned right in the middle of the field and is equipped with an empty backpack.

The motivation machine is implemented in Adam and functions as described in chapter 4 that runs within the Malmo mission loop. This means that his internal processes start running the moment he gets spawned into the world. Via an API in the motivation machine, we can request a dictionary of of dictionaries of both needs an consumptions and all its values. This way we can monitor the internal states of the agent, for example its need values, the need urgencies, the decay factor and so fort.[5]

Here we see the first impression when Adam gets spawned into the world. I have him equipped with a sword and in his inventory an apple. These items can be changed in the XML-file. The red harts symbolize its health in Minecraft and the drumsticks its food need in Minecraft.

[5]see the Appendix for the code



Synchronizing internal and external processes: time

One of the big challenges is synchronizing the internal processes of the agent that are represented in the motivation machine with the external processes of the outside world that is represented in Malmo's Minecraft. This challenge culminates where we have to update the agent needs. Updating the needs is a tricky part because this involves multiple time cycles since time is relative to world and agent. On the one hand we have the time of the motivation machine: the time for instance that it takes for a need to deplete or how long it takes before the consumption of an apple satisfies the food need. This time is dependent on the values of the variables in the motivation machine as explained in chapter 4. For example, when the gains, weight, loss and decay factor of the food need are high, the agent has a 'high' and thus fast metabolism.

On the other hand we have Minecraft time. Time in Minecraft comes in a day-night cycle of a 20 minute long lapse between two main light settings. But also the health and hunger of the agent have their internal Minecraft logic: Adam's health bar starts at 100% and slowly decay's and by default is connected to the hunger bar that is also shown on the heads-up display and represented by ten drumsticks. By the default the hunger bar decreases by players actions such as walking, sprinting, digging or attacking.

Tweaking all the variables independently of either the motivation machine or Minecraft to make them synchronized would be bothersome and makes the motivation machine too specific for one environment and thus inflexible. Therefore, I let the motivation machine have its own dynamic but regulate it with an extra loop on top of it as it advances in single steps that are independent from the time in the world. These steps update the motivation machine per frames/s in a separate configuration file. It works like a gas pedal: it makes the motivation machine in its entirety run faster or slower. [6]

But just as our biological clock has evolved in relation to night and day, we want the agent's clock to be related to the night and day cycle in Minecraft. This is important because one of the main reasons for implementing the motivation machine in a virtual agent was the relevance of situated cognition, of being in the world. To synchronize the motivation machine with the time in Minecraft, the steps of the motivation machine are updated with every *TimeStep* in Malmo: A counter that is related to the mission loop and increases simultaneously, that on its turn is connected to the time in Minecraft. This way the internal and external processes can be adjusted independently but also be related. This way a realistic interaction between agent and world is established.

4.3.3 What is working and what is not working?

Via the *AgentHost* we can give our motivational agent conditional commands, for example when its explore need value reaches a threshold, Adam starts walking and navigating the environment. By using observations from the *GetObservations* module in Malmo, Adam can identify objects in his environment and act accordingly. Right now, these decisions and observations are still hard coded, but further implementations could use computer vision to give motivated agents vision (Meulenbelt 2018) and could

[6] see Appendix

add learning to the motivation machine. Malmo is very suitable for reinforcement learning experiments. This way the agent can base its actions on (past) experiences.

In the current set-up, if Adam encounters an apple, he grabs it from the ground he puts it in his backpack. Based on the urge of the food need. he can decide to consume the apple. This triggers the 'eat' consumption. When he already has apples in his backpack however, he can decide to leave the apple but get for example bread - if he encounters that.

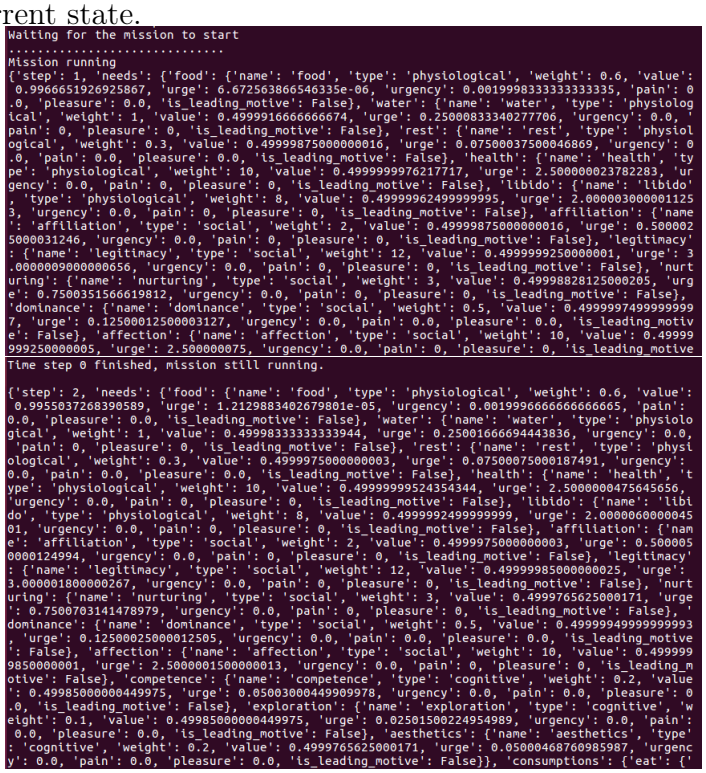
Navigation right now means that Adam randomly walks around within the boarder of the environmental grounds until the rest need makes him stop to activate the 'recover' consumption and deactivate the 'sprint' consumption. Except for quick recovery, we usually don't stand still in the middle of field to recover of course but it shows that Adam performs reasonable behavior solely on motivation given the conditions that are given by the environment.

Via the motivation machine API we can request the agent need and consumption states by which we can monitor the agent status while it is alive - that is: when a Malmo mission is running. Every *TimeStep* an update is sent to the researcher that shows the values of the internal state of the agent.

Here we see the output of the 2 time steps. It starts starts at zero.[7] The needs are in its initial state. The internal state of Adam changes as time progresses by updating

[7]the step of the motivation machine starts at 1

the needs. We can see Adam's needs change and each time step we can monitor his current state.



The researcher can now analyze this data and therewith he or she gets access to the internal state of the agent. This makes it possible to perform experiments in which the researcher formulates expectations and compares them with the observed behavioral data - as is common in many cognitive modeling experimental set ups.

4.3.4 What should be done to improve? And what can be expected?

The motivation machine has been developed independently from Minecraft. Although Minecraft in combination with Malmo gives us many options to modify the environment what makes it suitable as virtual world to experiment, it also means that we have to deal with limitations Minecraft and Malmo give us. Therefore the greatest challenge right now is to further synchronize the workings of the motivation machine with the workings of Minecraft.

Foods for example are consumable items. When eaten, they restore hunger. They are essential to survival and without the consumption of it, the agent will starve, causing damage to its health. Foods are eaten by 'using' them that can figure as a consumption event. However, the visual representation of the health bar in Minecraft can not be connected to the health need in the motivation machine and there is also no hack for that in Malmo[8]. So although theoretically possible, we are dependent on the progress the Malmo platform makes.

Similarly, a Minecraft agent doesn't need water but our motivated agent does. We can connect this to other consumable items, like milk. But again this doesn't work as intuitive as it should in an ideal virtual world. The same applies for the other physiological need *libido* since there are no sexual relationships possible in Minecraft.

There are opportunities for the social needs as Malmo just released a multi agents framework called MarLö that builds on top of the current Malmo framework to propose multi-agent and multi-task experimentation.[9] This could give the environmental conditions to involve *affiliation, legitimacy, nurturing, dominance, affection* into the mix that are ignored in the current experimental set up. The cognitive needs on the other hand can already be implemented in the current state.

Fortunately these are not insurmountable limitations and they are all theoretically possible to implement as Malmo further develops. Therewith, we now have a blueprint for further developing a computational theory of motivation that has been implemented in an agent in a virtual world.

[8]I have requested this many times but without success

[9]see: https://github.com/crowdAI/marLo

Chapter 5

Conclusion and Discussion

A theory of motivation should and can be part of a theory of the human mind. This theory should incorporate the different properties that have been argued about over the years, like learning strategies, innateness, different categories of needs, hierarchy of needs (what needs should be satisfied first) and how much such a need based system can predict parts of behavior. This way we can answer why organisms move in the first place, predict it and thus model it.

To explore the requisites of a computational model of motivation, I considered various theories of motivation. Sigmund Freud's theory of drives contained much of the main ideas that were made concrete later by others. He introduced the idea of innate needs, that translate into drives, wishes or desires. From there on researchers have pulled motivation theory into the sciences by giving it predictive properties and were able to proof their hypotheses. An important finding was that the function of most of our actions is to serve the needs. Motivation causes this activity and activity has many forms, like dreams, aspirations, thoughts, feelings and indeed behavior. Furthermore, recent studies in motivation confirmed the universal existence of needs and their psychological and social characteristics.

Artificial Intelligence is a good place to develop and test such a motivational theory. It can verify different hypotheses by designing different configurations and test them by building them. Such an attempt has been made by the cognitive architecture MicroPsi. A cognitive architecture that tries to model the human mind. MicroPsi's core and distinctive characteristic is motivation. However, exactly this feature hasn't been implemented yet and is still under development. To evaluate MicroPsi's motivational core, I explored and described its inner workings and configurations. The motivation machine meets most conditions, although it still has a very passive image of organisms and relies on the the old idea of homeostasis. But the only way to really test the motivation machine's underlying theory and further develop it, is to implement it into an agent and put it into the world.

This could be done into a real world, but I believe that a virtual world meets the minimal criteria. Therefore I have implemented the motivation machine into Minecraft using the Malmo API. Therewith I have build a motivational agent into a virtual world. This agent is now exposed to a large variety of tasks, challenges and situation and the researcher has experimental control. This provides a minimal working model of motivation, a blueprint for further motivation research.

Concretely it means that a motivational model is running on a Minecraft-agent. The initiated agent - called Adam - gets spawned into the world. By using the *AgentCommand* he can be given a varied amount of tasks. The need values are printed out and are thus available for analysis. This way the experimenter can get insight into the internal state of the agent, something that has been lacking in all other motivational research. The other needs run as well according to the configuration in the motivation machine, but they are currently not in interaction with the virtual world.

There are some limitation to my thesis that creates opportunities for further research. When God created life, he had also created the universe. When I created

Adam, I had to use an existing universe. Since any world has its own (hard-coded) logic, our virtual world - or real world if you implement the motivation machine into a robot - has synchronizing problems too. One of the best universes available right now is Minecraft due to its flexibility, open character and the Malmo API. Unfortunately the combination of the motivation machine and Minecraft has many synchronizing problems still. An important problem that I have solved is time. Time in Minecraft and time in the motivation machine is now in harmony and controllable. But other problems that cry for a solution are metabolism (how long does it take to digest an apple?), sleep cycles, hydration (there is no water in Minecraft), interaction with other agents (to satisfy social needs) to name a few. Also health in Minecraft and health in the motivation machine could be synchronized. Not only for experimentation purposes; it would also be visually attractive to have the health bar correspond with the underlying mechanism.

These problems are not insurmountable, they have to be solved by gradual experimentation or by hacks. We could for example replace water with an item that is available in Minecraft like milk. But for a durable working model, we need to edit the motivation machine and the virtual world together until they both function in logical interaction.

The next step is to further compare the current configuration of the motivation machine with the motivation theories as I have described them thoroughly in this thesis. I am sure that there are more specific theories that give info about the specific configuration of the motivation machine. This info can lead to new hypothesis and - if turned into code - can immediately be implemented as long as Malmo allows for it.

This is not something that should be done at once. The whole idea of having a computational theory of motivation is that one can have an hypothesis of how motivation works, implement these in a way a computer can understand it, and test the hypothesis interactively. Based on the given the results, the researcher can get

back to his theory, change its parameters and run it again until the behavior of the agent is similar to the predictions of the theory.

Another limitation of the current implementation is the lack of other cognitive capacities. My current agent lacks memory and learning strategies. If Adam finds an apple and walks back to his starting point to rest, it will not remember where it found the apple and whether eating it solved his hunger or made him food poisoned. In motivation machine language: Adam right now has no competence. The current design of the motivation machine is unable to solve this as I've shown in chapter 3. Subsequent research could involve reinforcement learning that trains the agent and therewith makes it a more competent in serving its needs. The current set-up offers plenty of opportunity for such an approach.

An approach I would recommend to solve the lack of other cognitive capacities, like vision, memory and navigation is intergrating the motivation machine to a cognitive architecture that has these modules developed and implemented already like MicroPsi. A connection between Minecraft and MicroPsi currently exists due to the work of Dirk Meulenbelt (Meulenbelt 2018). But I speak from experience if I say that connecting anything to MicroPsi involves many difficult to solve dependencies and a steep learning curve to fully comprehend the workings of the architecture.

Lastly I would like to mention the problem of goal alignment in AI. The motivation machine in its current state has the potential to let the agent create, prioritize and identify its own goals, within the range of possible needs that we endowed it with. Subsequent research could fulfill this potential. Every event has the option to become a goal if it is associated with a leading motive. This means that the agent has to learn what consumption has motivation relevance and accordingly what actions should follow. Learning techniques could train the agent to find the corresponding action in Minecraft to reach that goal. This would be a first step towards an important challenge in AI: trying to align goals from the bottom up instead of top-down.

With this thesis I have followed the old advice of Herbert Simon from 1967. If we want to understand the workings of the mind, if we want to achieve Artificial General Intelligence, we need to investigate the rich field of motivation research and integrate motivation into our AI-models. Motivated agents are now situated within an excellent world for AGI experimentation. It provides a proof of concept for motivated agents in virtual words and gives way to further experimentation.

References

Anderson, J.: 2009, *How Can the Human Mind Occur in the Physical Universe?*, Oxford University Press, Oxford.

Atkinson, J.: 1958, *Motives in Fantasy, Action and Society*, Van Nostrand, Princeton.

Bach, J.: 2007, *Princilples of Synthetic Intelligence: Building Blocks for an Architecture of Motivated Cognition.*, PhD thesis, University of Osnabrück, Ösnabruck.

Bach, J.: 2012, MicroPsi 2: The next generation of the MicroPsi framework, *International Conference on Artificial General Intelligence*, Springer, pp. 11–20.

Bach, J.: 2017, An Adaptable Version of the MicroPsi Emotion Model.

Boden, M. A.: 2016, *AI: Its Nature and Future*, first edition edn, Oxford University Press, Oxford, United Kingdom.

Bostrom, N.: 2016, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford, United Kingdom ; New York, NY. OCLC: ocn945184787.

Brooks, R. A.: 1991, Intelligence Without Reason, *IJCAI-91* p. 28.

Cassirer, E.: 1944, *An Essay on Man*, New Haven: Yale University Press.

Chomsky, N.: 1971, The Case Against B.F. Skinner, *The New York Review of Books* p. 12.

Darwin, C.: 2011, *On the Origin of Species by Means of Natural Selection: Or the Preservation of Favoured Races in the Struggle for Life*, ProQuest LLC ; Penguin, Cambridge; London. OCLC: 803439483.

Deci, E. L. and Ryan, R. M.: 2000, The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior, *Psychological Inquiry* **11**(4), 227–268.

Dennett, D. C.: 1987, *The Intentional Stance*, MIT Press, Cambridge, Mass.

Dennett, D. C.: 2018, *Van bacterie naar Bach en terug: de evolutie van de geest.* OCLC: 1021829845.

Dörner, D.: 1999, *Bauplan Für Eine Seele.*, Reinbeck.

Dörner, D. and Güss, C. D.: 2013, PSI: A computational architecture of cognition, motivation, and emotion., *Review of General Psychology* **17**(3), 297–317.

Feyerabend, P.: 1975, *Against Method*, New Left Books, London.

Freud, S.: 1915b, Instincts and their Vicissitudes, *in* I. Smith (ed.), *Freud Complete Works*, 2011 edn.

Goertzel, B. and Pennachin, C. (eds): 2007, *Artificial General Intelligence*, Cognitive Technologies, Springer, Berlin ; New York.

Hadfield-Menell, D., Dragan, A., Abbeel, P. and Russell, S.: 2016, The Off-Switch Game, *arXiv:1611.08219 [cs]* .

Hull, C.: 1943, *Principles of Behaviour*, Yale University Press, New York.

Johnson, M., Hofmann, K., Hutton, T. and Bignell, D.: 2016, The malmo platform for artificial intelligence experimentation, *International Joint Conference on Artificial Intelligence (IJCAI)*, p. 4246.

Kant, I.: 1963, *Kritik Der Urteilskraft*, Reclam. 00000.

Kant, I.: 1987, *Kritik Der Reinen Vernunft*, Hamburg: Felix Meiner Verlag. 07972.

Kotseruba, I., Gonzalez, O. J. A. and Tsotsos, J. K.: 2016, A Review of 40 Years of Cognitive Architecture Research: Focus on Perception, Attention, Learning and Applications, *arXiv preprint arXiv:1610.08602* .

Leibniz, G.: 1714, The Principles of Philosophy known as Monadology.

Maslow, A. H.: 1943, A Theory of Human Motivation, *Psychological Review* (50), 370–396.

Maslow, A. H.: 1987, *Motivation and Personality*, 3rd ed edn, Harper and Row, New York.

McClelland, D.: 1985, *Human Motivation*, Scott Foresman, Glenfield.

McDowell, J.: 1996, *Mind and World*, Cambridge, MA: Harvard University. 00000.

Meulenbelt, D.: 2018, *Sighted MicroPsi Agents in Minecraft: Object Recognition Using Neural Transfer Learning and Automated Dataset Collection*, PhD thesis, Utrecht University, Utrecht.

Minsky, M.: 1986, *The Society of Mind*, Simon and Schuster, New York.

Minsky, M.: 2007, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind*, Simon & Schuster, New York. OCLC: 254260834.

Misselhorn, C.: 2009, Empathy with Inanimate Objects and the Uncanny Valley, *Minds and Machines* **19**(3), 345–359.

Murray, H.: 1938, *Explorations in Personality*, Oxford University Press, New York.

Newell, A.: 1990, *Unified Theories of Cognition*, number 1987 in *The William James Lectures*, Harvard University Press, Cambridge, Mass.

Pinker, S.: 1997, *How the Mind Works*, Norton, New York.

Pintrich, P. R.: 2003, A Motivational Science Perspective on the Role of Student Motivation in Learning and Teaching Contexts., *Journal of Educational Psychology* **95**(4), 667–686.

Quine, W. V.: 1951, Main Trends in Recent Philosophy: Two Dogmas of Empiricism, *The Philosophical Review* **60**(1), 20–43. 06206.

Reeve, J.: 2009, *Understanding Motivation and Emotion*, 5th ed edn, John Wiley & Sons, Hoboken, NJ. OCLC: ocn230191075.

Russell, B.: 1912, *The Problems of Philosophy*, Digireads.com Publishing, Lawrence, KS.

Russell, S., Dewey, D. and Tegmark, M.: 2015, Research Priorities for Robust and Beneficial Artificial Intelligence, *AI Magazine* **36**(4), 105.

Searle, J. R.: 1980, MINDS, BRAINS, AND PROGRAMS, *Behavioral and Brain Sciences* **3**(3), 417–457.

Shannon, C. and Weaver, W.: 1964, *The Mathematical Theory of Communication*, The university of illinois press, Urbana.

Simon, H. A.: 1967, Motivational and emotional controls of cognition., *Psychological Review* **74**(1), 29–39.

Skinner, B. F.: 2002, *Beyond Freedom & Dignity*, Hackett Pub, Indianapolis, Ind.

Sloman, A.: 1981, Why robots will have emotions, *Why Robots Will Have Emotions*, Vancouver.

Tay, L. and Diener, E.: 2011, Needs and subjective well-being around the world., *Journal of Personality and Social Psychology* **101**(2), 354–365.

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

Villeneuve, D.: 2016, Arrival. IMDb ID: tt2543164.

Winter, D.: 1973, *The Power Motive*, Free Press, New York.

Appendix

Here are the python files that connect the motivation machine with Minecraft, using the Malmo-API, create a mission and start the simulation. I have include the three essential files here. The XML settings set up the experimental environment. The simulation.py runs the actual simulation experiment. The whole repository can be found on my github: https://github.com/reiniertromp. This is a better place to copy the code because of indentation (latex has difficulties with line breaks). I would advice to keep the appelboer.py and the XML settings separate.

Appelboer.py

```
from __future__ import print_function
from builtins import range
import MalmoPython
import os
import sys
import time
import json
from xml_settings import *
#import agent
import threading


# optional imports
from model.needs import needs, consumptions
from model.agent import Need, Modulator, Consumption
from random import randint, random
import math
```

```
from widgets import Settings
from helper_widgets import Diagram

from model import api
from model.needs import needs, consumptions
from model.modulators import modulators, aggregates
from model.emotions import emotions
from simulation import Simulation

if sys.version_info[0] == 2:
    sys.stdout = os.fdopen(sys.stdout.fileno(), 'w', 0)  # flush print o
else:
    import functools
    print = functools.partial(print, flush=True)

# Create default Malmo objects:
class Adam:

def __init__(self):
    self.moving = False
    self.resting = False
    self.mission_started = False
    self.data = []
    self.time_step = 0
    self.inner_Adam = Simulation()
    self.agent_host = MalmoPython.AgentHost()
    self.reset = api.reset()
```

```
    try:
        self.agent_host.parse( sys.argv )
    except RuntimeError as e:
        print('ERROR:',e)
        print(self.agent_host.getUsage())
        exit(1)
    if self.agent_host.receivedArgument("help"):
        print(self.agent_host.getUsage())
        exit(0)

#Build a mission and give it the configuration of the world.
Note that we can change that later with Python (not XML) code.
self.my_mission = MalmoPython.MissionSpec(missionXML, True)
self.my_mission_record = MalmoPython.MissionRecordSpec()

#We may also store previous observations within Malmo,
but let MicroPsi take care of remembering what they were.
self.agent_host.setObservationsPolicy
(MalmoPython.ObservationsPolicy.LATEST_OBSERVATION_ONLY)
self.my_mission.forceWorldReset()

def start_mission(self):
    max_retries = 3
    for retry in range(max_retries):
        try:
            self.agent_host.startMission(self.my_mission, self.my_missio
            break
```

```
        except RuntimeError as e:
            if retry == max_retries - 1:
                print("Error starting mission:",e)
                exit(1)
            else:
                time.sleep(1)

def loop_until_mission_starts(self):
    print("Waiting for the mission to start ",)
    self.world_state = self.agent_host.getWorldState()
    while not self.world_state.has_mission_begun:
        sys.stdout.write(".")
        time.sleep(0.1)
        self.world_state = self.agent_host.getWorldState()
        for error in self.world_state.errors:
            print("Error:",error.text)
    print()
    print("Mission running ",)

def get_world_state(self):
    #Helper function to ensure we get the world state.
    world_state = agent_host.peekWorldState()
    while world_state.is_mission_running and all (e.text=='{}' for e in
        world_state = agent_host.peekWorldState()
    return world_state
```

```
def get_observations(self):
    world_state = self.get_world_state()
    if not all(e.text=='{}' for e in world_state.observations):
        obs = json.loads( world_state.observations[-1].text )
    else:
        raise Exception('\nSomehow did not get observations, despite wai
    return obs

def get_rest_values(self, data):
    """returns a list of values of the rest need  """
    rest_values = [step["needs"]["rest"]["value"] for step in data]
    return rest_values[-1]

def appelboer(self):
    if not self.mission_started:
        self.start_mission()
        self.loop_until_mission_starts()
        self.mission_started = True

    #Check whether the mission is actually running. (When MicroPsi think
    else:
        world_state = self.agent_host.peekWorldState()
        if not world_state.is_mission_running:
            raise Exception('\nMission no longer running!\n')


    if self.time_step == 10:
```

```
            self.agent_host.sendCommand('move 1')
            self.moving=True

        print(api.get_needs_and_consumptions())

    # optional for consumption triggers
    # print(api.get_needs())
        #
        # if self.moving:  # agent_host.sendCommand("move 1"):
# IF ADAM MOVES TRIGGER SPRINT CONSUMPTION
        #
        #     consumptions["sprint"].trigger()
        #     consumptions.update()
        # elif not self.moving:  # agent_host.sendCommand("move 0"):
# IF HE STANDS STILL, TRIGGER RECOVER CONSUMPTION
        #     consumptions["recover"].trigger()
        #
        # self.data = self.inner_Adam.log
        # rest_values = self.get_rest_values(self.data)
        # print(rest_values)

        print('\nTime step {0} finished, mission still running.\n'.format(st
        self.time_step += 1

    def start_simulation(self):
        self.my_mission.forceWorldReset()
        print("Start nu de main loop.")
```

```
        while self.inner_Adam.step():
            self.appelboer()
        print("\nAfgelopen.\n")

if __name__ == "__main__":
reinier = Adam()
reinier.start_simulation()


XML settings:
from object_constructors import *

missionXML ='''<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<Mission xmlns="http://ProjectMalmo.microsoft.com" xmlns:xsi="http://www

<About>
<Summary>Appeltjes!</Summary>
</About>

<ServerSection>
<ServerHandlers>
<FlatWorldGenerator generatorString="3;7,2*3,2;1"/>
<DrawingDecorator>

<!-- Walls around the experimental ground -->
<DrawCuboid x1="-50" x2="50" z1="-50" z2="-50" y1="4" y2="100" type="obs
<DrawCuboid x1="-50" x2="50" z1="50" z2="50" y1="4" y2="100" type="obsid
```

```
<DrawCuboid x1="-50" x2="-50" z1="50" z2="-50" y1="4" y2="100" type="obs
<DrawCuboid x1="50"  x2="50" z1="-50" z2="50" y1="4" y2="100" type="obsi

<!-- Add trees / food -->
''' + gen_lots_of_food() + '''

</DrawingDecorator>
</ServerHandlers>
</ServerSection>

<AgentSection mode="Survival">
<Name>Adam</Name>

<AgentStart>

<Placement x="0" y="4" z="0" yaw="0"/>

<Inventory>
<InventoryItem slot="0" type="diamond_sword"/>
<InventoryItem slot="1" type="apple"/>
</Inventory>

</AgentStart>

<AgentHandlers>
<MissionQuitCommands/>
<ObservationFromFullStats/>
```

```
<!--      <ObservationFromFullInventory/>  -->
<ContinuousMovementCommands turnSpeedDegs="180"/>
</AgentHandlers>
</AgentSection>
</Mission>'''


Simulation.py
from widgets import Settings
from helper_widgets import Diagram

import time
from model import api
from model.needs import needs, consumptions
from model.modulators import modulators, aggregates
from model.emotions import emotions


class Simulation(object):
def __init__(self):

    api.reset()
    self.needs = list(needs.values())
    self.consumptions = list(consumptions.values())
    self.modulators = list(modulators.values())
    self.aggregates = list(aggregates.values())
    self.emotions = list(emotions.values())
```

```
        self.need_index = needs

        self.current_simstep = 0

        self.log = []

    def step(self):
        """Advances the simulation by a single step. Returns False if we are
        if self.current_simstep < Settings.max_simulation_steps:
            # for consumption in self.consumptions:
            #     if random()>0.99:
            #         consumption.trigger()
            time.sleep(5)
            api.update()
            self.current_simstep += 1
            self._update_log()
            return True
        return False

    # def step(self):
    #     """Advances the simulation by a single step. Returns False if we a
    #     if self.current_simstep < Settings.max_simulation_steps:
    #         consumptions['eat'].trigger()
    #         api.update()
    #         self.current_simstep += 1
    #         self._update_log()
    #         return True
```

```
#     return False

def _update_log(self):
    """adds the current values to the log."""
    self.log.append(api.get_data())
```