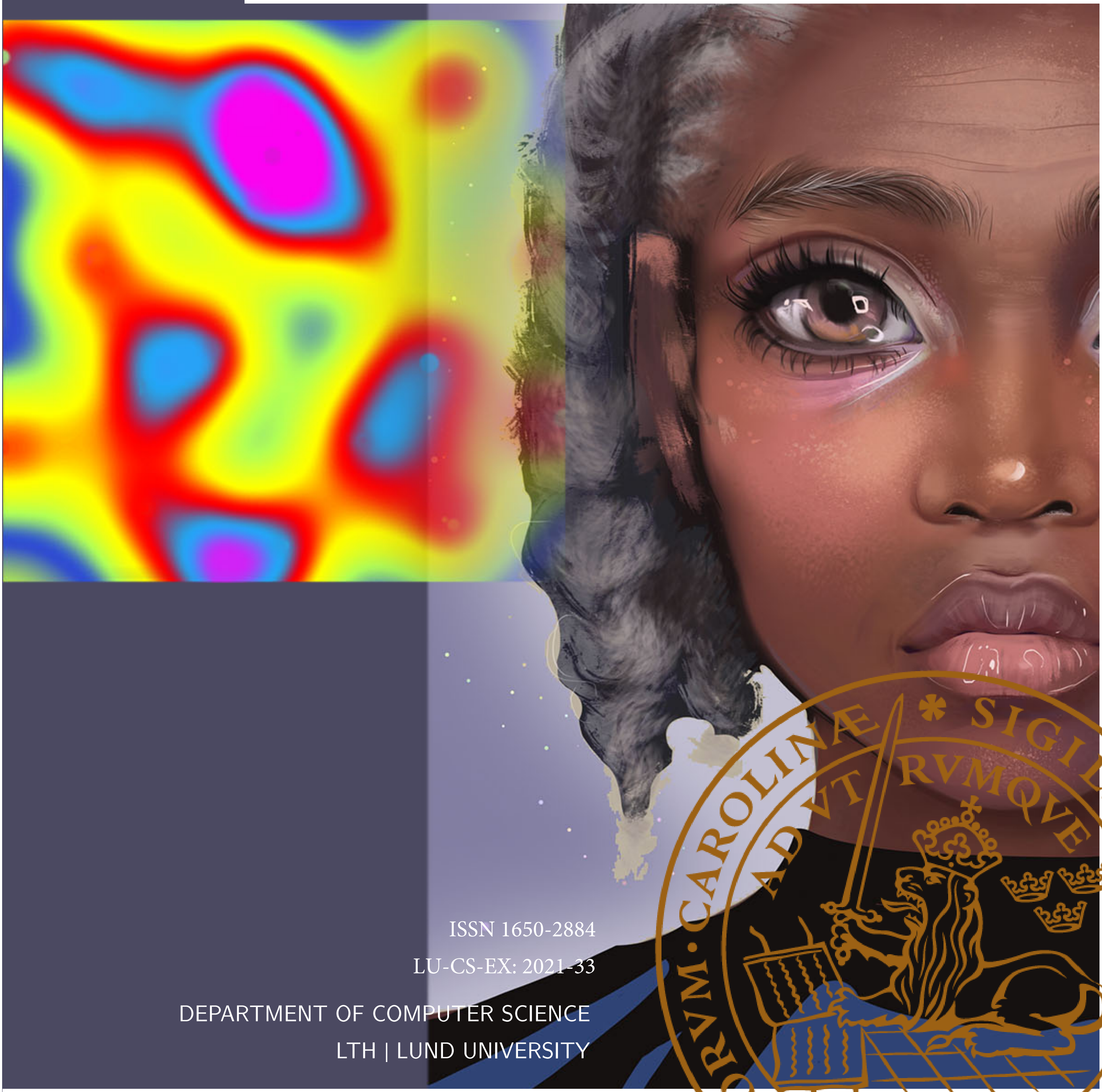
BACHELOR'S THESIS 2021

Motivating a Conscious Machine with Novelty-Seeking

Miroslava Brodlova



ISSN 1650-2884
LU-CS-EX: 2021-33

DEPARTMENT OF COMPUTER SCIENCE
LTH | LUND UNIVERSITY

KANDIDATARBETE
Datavetenskap

LU-CS-EX: 2021-33

Motivating a Conscious Machine with Novelty-Seeking

Att Motivera en Medveten Maskin med Nysökande

Miroslava Brodlova

Motivating a Conscious Machine with Novelty-Seeking

Miroslava Brodlova
`miro.brodlova.420@student.lu.se`

June 3, 2021

Bachelor's thesis work carried out at
the Department of Computer Science, Lund University.

Supervisor: Elin Anna Topp, `Elin_Anna.Topp@cs.lth.se`

Examiner: Jacek Malec, `Jacek.Malec@cs.lth.se`

Abstract

This bachelor's thesis aims to be a starting point for constructing motivation in a conscious machine through a novelty-seeking mechanism. Firstly, a limited literary review on how motivation functions in humans and animals is conducted to extract clues on how motivation may be implemented in a conscious program. Secondly, based on Lenore and Manuel Blum's model of a Conscious Turing Machine (CTM), basic components that theoretically support consciousness in non-biological matter are built. The CTM parts constructed include a Long Term Memory (LTM), a Short Term Memory (STM), Chunks and an Up- and Down-stream. Visual stimulus is then sent into the CTM-program in the form of pixel-values. Receiving pixels-values that are unchanging are in this project classified as a negative experience for the CTM, whereas pixel-values that vary over time are positive. The results highlight the difficulty in building a conscious machine with concurrent visual input. The project's resulting program was only able to effectively handle about 60000-250000 LTM-processors (a far cry from the human brain's 86 billion neurons). More research is suggested that focuses on how the CTM can interact with visual stimuli and novelty-seeking algorithms for the purpose of finally achieving machine consciousness.

Keywords: Conscious Turing Machine, CTM, Novelty-Seeking, Boredom, Conscious AI, Artificial Intelligence

Acknowledgements

I would like to thank my Supervisor Elin Anna Topp and my Examinator Jacek Malek that have taken on helping me with this project despite their very busy schedules.

I would also like to greatly thank Lenore Blum and Manuel Blum who went above and beyond in helping me understand their model.

Thanks also goes to the Jean-Louis Villecroze, Marcus Klang and the students at Peking University for generously sharing their knowledge, work and classroom with me so that I could explore the CTM further.

Contents

Chapter 1

Introduction

This Bachelor's thesis aims to investigate the construction of artificial consciousness and focuses on building basic, core components of a program that may eventually achieve awareness. Due to the magnitude of the task of building machine consciousness, this report places its focal point on one aspect necessary for awareness, namely inherent motivation, as it has yet not been explored in greater detail [9, 10]. The report is split up into five sections; *Introduction, Theory, Implementation, Result* and *Evaluation*.

Firstly, the background to this thesis is detailed in the Introduction section. Later, how one may construct inherent motivation in a conscious machine, in a way that corresponds with how motivation functions in humans and nonhuman animals, is examined in the Theory section of this report. A prototype of how this functionality can be achieved in a machine is also proposed. In the Implementation section the actual construction of the parts of the conscious model built in this project are described in greater detail. Finally, in the Results and Evalutation sections the consequent program is scrutinized. Focus is also placed on to what degree the program built in can be considered conscious.

1.1 Previous research

Ample research exists that reviews the ethical and philosophical aspects of creating machine consciousness [40, 1, 51]. However, few articles describe practical aspects or actual attempts at construction of machine awareness. Conceivably, this is because approachable and concrete models of artificial consciousness have so far been unavailable. However, in recent years models based on current neurological, psychological and computer-based science have emerged [42, 22, 15, 9]. These provide concrete definitions of machine consciousness that aim to be compatible with what we may achieve in computers today or in the not too distant future.

1.2 Blum's and Blum's model

In March 2021, after 65 years of cumulative research [9], one such model christened the Conscious Turing Machine (CTM) was presented by computer scientists Lenore and Manuel Blum. Their model details a hyper-simplified model of human consciousness based on interconnections between a Short Term Memory (STM) and Long Term Memory (LTM) processors. Information from the LTM can bubble up to the STM and depending on its content causes the machine pain or pleasure. In their article they argue that the emerging interconnectedness between the memories, and the pain and pleasure content that can make its way to the STM, will result in a sort of machine awareness and experience[9]. The Blums' model of consciousness is architectural, meaning that it is the structure and dynamics of the components of the CTM that they argue create cognition. Their explanation of consciousness therefore works as well for machines with brains of gold and silicon as with animal brains made from flesh and blood [9]. Furthermore, although there are several models of artificial consciousness just beginning to emerge [22, 47], the CTM is arguably currently the most formalized. This makes the model especially exciting and suitable for further exploring artificial consciousness. Important to note is that although parts of the CTM are described in great detail by the Blums, more research needs to be done on how to further formalize vital components of the CTM for the model to finally be a complete blueprint on how to build computer consciousness [9].

1.3 Goal

The Blums argue that interdisciplinary research indicates that there are three categorically essential components that need to be included in a machine for it to achieve consciousness; an inner voice, the ability to think and motivation [9, 10]. The first two aspects, the inner voice and the ability to think, have been detailed in greater depth in their paper. While more research still needs to be done on how to actually construct these components, they are well-defined in their own right. In contrast, manufacturing motivation in the CTM has not been as thoroughly explored. This paper aims to mitigate this by investigating, in part through reviewing literature on motivation and in part through action research, how motivation can be constructed in the CTM. Ultimately the main contribution of this Bachelors' thesis is its attempt at formalizing one of the simplest ways motivation may be constructed in the CTM and then begin building a prototype of this using the CTM model.

1.4 Risks

There are substantial risks in attempting to build programs and components that ultimately aim to capture consciousness. Research fields that deal primarily with conscious subjects, like psychology and medicine, often have reprehensible pasts because of their dismissiveness of the harm their experiments could cause [12]. For instance, in 1966 psychologist John Money, convinced that people were born gender neutral, caused one boy in a set of twins to undergo sex reassignment surgery (turning his genitals into a vulva). Money continued to have meetings with the twins over 10 years, sometimes encouraging the children to perform sexual acts together. Neither of the boys ended up ever identifying as a female growing up and possibly partly due to the trauma Money inflicted, both brothers committed suicide as adults[20]. Humans are additionally frighteningly adept at dismissing the conscious experience of others. Before 1999 it was believed that babies, despite visible crying, were not able to feel pain [3]. Unsurprisingly, that has been proven to be false. Infants do feel intense pain, and research indicates that they probably experience it more intensely than adults do [44].

These aspects are important to mention, as they hint at our ability to dismiss painful and conscious experiences of others, even when they are expressed via verbal screams within our own species. There is therefore a real risk that we will not recognize, or even simply dismiss, consciousness and pain emerging from a machine when we have created it. Subsequently, we have a moral and ethical responsibility, no matter how far away from succeeding we think we may be, to always consider these questions when attempting to create awareness where pain is involved.

For this reason this thesis will also include efforts to mitigatate possible agonizing experiences in the CTM and attempt to find ways in which a machine can turn off itself if the "pain"-threshold becomes high. At this stage taking such measures may be predominantly symbolic but is nevertheless important.

1.5 Research Questions

1. How can rudimentary motivation be manufactured in the CTM?

2. To what degree do the CTM-components constructed in this project facilitate motivation?

3. To what degree can this project's resulting program, based on the CTM model, be considered conscious?

1.6 Contributions

Throughout this report assistance has been given in constructing the CTM components in an accurate and efficient manner. Manuel and Lenore Blum gave feedback on first drafts of this report, assisting in raising the thesis' quality. The pseudocode contained in their 2021 paper [9] was also used as a basis for building several of the CTM components in this project. Furthermore, Jean-Louis Villecrose, a software engineer who is currently developing a program that aims to capture language comprehension in the CTM, generously shared executable versions of his project via email correspondence to assist in this thesis. He also shared the documentation of one of his earlier CTM projects [53]. Although no source code from his scripts was used in this paper, it was helpful to see how his programs worked. Finally, Marcus Klang spent over an hour assisting in this project via video-call when the program displayed performance issues. He gave guidance on what aspects may be tweaked to yield a more efficient program.

Therefore, although no actual code has been used in this thesis that originates from somewhere else, much generous help was given to make this paper possible.

Chapter 2

Approach

2.1 Method

In the first part of this project, with the purpose of finding an adequate motivational mechanism for the CTM, a limited literary review on human and nonhuman animal motivation was conducted. Research databases such as LubSearch and Google Scholar were used to find research papers on motivational mechanisms in animals as they are understood today. Keywords used to search for scientific papers on the topic included "motivation", "brain", "animal", “human” and later on "novelty-seeking". Due to the limited scope of this Bachelor’s thesis (only 10 weeks) only around 7 papers were investigated in closer detail and were then used to produce a suggested structure for how motivation can be built in the CTM, as described in the Theory section below.

For the second part of the project a prototype of the proposed motivational structure in CTM-components was built. To more easily facilitate the construction of the CTM components the program was built in Unity, a development platform often used for game-creation. The program was built on a home computer and the installed processor was Intel Core i5-4460. The GPU in use was NVIDUA GeForce GTX 1060 3GB. The details of the subsequent program are described in the Implementation section below.

2.2 Theory

Below some vital CTM-components are described briefly to give context to the work done in this thesis. The content of the literary review relating to motivation in humans and nonhuman animals is also presented. Finally the theoretical cross-section between motivation and the CTM is described. Hence, three categories are described in detail below: *The components that build the CTM*, *Motivation in human and nonhuman animals* and lastly *Motivation and the CTM*.

2.2.1 The components that build the CTM

The CTM is composed of seven overarching components; the Short Term Memory (STM), the Long Term Memory (LTM), the Up-Tree, the Down-Tree, Links, Input and Output (see Figure 2.1) [9]. The unconscious LTM-processors are run by algorithms that produce information, or *thoughts*. They put these thoughts in data-packets called Chunks every time-unit t. Each Chunk is assigned a weight to indicate how important the processor computes that that thought is. Every produced Chunk at time t is then part of a competition via the Up-Tree to reach the conscious STM. The winning Chunk becomes the thought the CTM is thinking. Every t the STM also broadcasts to all LTM-processors which Chunk, or thought, it currently has. This way, the unconscious processors are always aware of what the STM is "thinking", and can incorporate that information into their calculations in case the thought has relevance to their algorithms. It is partly through this constant flow of information between the STM and LTM, and links emerging between processors, that consciousness arises. It is important to highlight that this description of the CTM is extremely simplified, and to fully understand the CTM in detail and why Manuel Blum and Lenore Blum argue it is conscious one should read their 2021 paper [9].

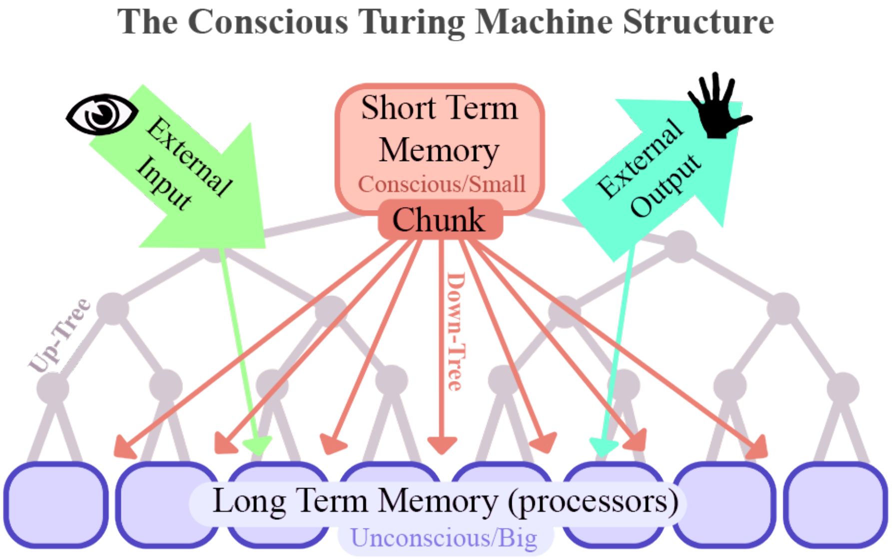


Figure 2.1: A simplified illustration of the CTM model based on the Blums' design presented in their 2021 research paper [9]. The Down-Tree is illustrated through the arrows streaming from the Chunk under the Short Term Memory. The Up-Tree is illustrated as a binary tree emerging from the LTM. Links (not illustrated here) emerge between LTM-processors when they find useful information in one another.

2.2.2 Motivation in human and nonhuman animals

Motivation, referring to the physiological and psychological mechanism by which people choose particular behaviors and persist with them [36], wields powerful influence over our daily lives. Despite its integral role in human and nonhuman behavior however, studies that examine the neurological components of how motivation works have only recently started to emerge [16].

Motivation driven by pleasure seeking

Modern studies demonstrate that motivation is sub-served by dopaminergic systems, linking it to areas in the brain that deal with the sensation of reward and pleasure [54, 14]. This may seem unsurprising as there are numerous pleasurable states that clearly motivate humans and nonhuman animals. The desire to feel satiated, to feel warmth, to feel included in a social group and the desire to be recognized are examples of such enjoyable experiences. Typical to human complexity however, what we are motivated to do is not always evidently linked to pleasure. Sometimes humans are seemingly motivated to seek out their own immediate suffering and destruction. Self-harm and suicides are examples of this and are depressingly the second most common cause of death for young people worldwide [25]. Such behaviors appear to contradict the notion that motivation and behavioral drive are dependent on reward and pleasure. On this it is important to note that research indicates physical and psychological pain may be similar, and may also stand in conflict with each other. In 2011 a brain-imaging study subjected participants to social rejection by peers and remarkably found that the brain showed activity in the same areas as when people are experiencing physical pain, suggesting that emotional and physical pain may be more alike than previously thought [19]. Moreover, sufferers of mental disorders such as depression have been known to describe their psychological pain as worse than any physical pain they have ever experienced [37]. In such a state attempting to end one's life can be a drastic and immediate way to try to stop that pain. Individuals who have self-harmed have likewise expressed that the overwhelming emotional agony they felt dwarfed the physical pain they cause themselves. By injuring themselves they allowed their focus to shift on pain that was less distressing [34]. Even self-harming behaviors are therefore arguably motivated by the desire to increase contentedness and minimize pain.

Motivation driven by novelty seeking

As described above, emotional experiences like pleasure and pain intermingle in humans and create complexly motivated behaviors that are sometimes difficult to untangle. Beyond this human and nonhuman animals alike partake in behaviors where the connection to pleasure and pain is not as clear. Intrinsically motivated activities are examples of these. Such activities involve novelty seeking rather than pleasure seeking, and are behaviors that are directed towards the unknown and uncertain [35]. They display our tendency to yearn for novel stimuli [54] and may play an important role in our development as they motivate us to investigate the unknown and learn about the world [35].

In humans this tendency is evident in many ways. For instance, the human gaze is involuntarily attracted to objects in the environment that change (more so than objects that remain static) [27, 28]. Variety seeking also guides our conscious actions in our daily lives, such as when making commercial purchasing decisions [24]. Even infants display novelty-seeking behavior and show a preference for novel rather than familiar stimuli as they grow older [13]. Furthermore, this trait is not unique to humans. Research shows that rats, among other animals, prefer environments that have been paired with novel stimuli [5]. Prolonged lack of novel stimuli will even cause significant mental harm. For instance, rhesus monkeys that are housed alone for prolonged periods of time have been shown to display self-injurious behavior to cope [33], seemingly preferring painful stimuli to no stimuli at all. In humans the effects of prolonged sensory deprivation are so severe they can cause psychotic breaks, hallucination and depersonalization [49]. Lack of varying stimuli to humans is so damaging that solitary confinement and sensory deprivation classifies as torture [39, 38]. From this it is evident that human and nonhuman animals need a degree of shifting stimuli in our environment to thrive. Such novel stimulus acts as a motivational force pulling us towards it and invites our further exploration.

2.2.3 Motivation and the CTM

Due to the complex and messy nature of motivation in conscious humans and nonhuman animals it is a challenge to fit the concept into the simplified CTM. Nevertheless one approach may be to isolate one type of binary motivation, such as novelty seeking, and attempt to simplify it to its basic components so that it can be implementable in the model.

Novelty-seeking in the CTM

Novelty seeking in particular may be a suitable motivational mechanism for the CTM as it operates in only two dimensions. Varying stimuli motivate us to investigate, while sameness in stimuli does not elicit responsiveness. This can be contrasted with using physical pain as a motivator. As described above, physical pain can be something that motivates us to stop what we are doing, but it can also attract us towards it (such as in cases with self-harm). This paper makes the argument that novelty rewarding algorithms in the LTM processors, partnered with simple visual input data, may be one of the simplest forms of motivation that can be captured in a conscious AI.

How to concretely fit motivation in the CTM

One way to possibly elicit motivation in the CTM may be to first connect it to a digital, visual input. This visual input can be programmed to send pixel-data (such as color-information) to individual LTM-processors. The processors can then be equipped with algorithms to calculate weight based on the variety of pixel-data as the unit of time t is increasing. If the pixel-data remains similar negative weights will be calculated (simulating the averse human response to prolonged stimuli sameness) while positive weights are calculated if pixel-data changes (simulating human interest in novel stimuli). If the CTM was then given an actuator to navigate the screen with a mouse it could possibly navigate to screen-content that causes more positive weights to reach the STM. This design creates a basic playing field where motivation to find novel input may occur. It is this design that will be attempted to be built in this project, and is further described in the Implementation-section below.

2.3 Implementation

This section will detail how each CTM-component was constructed in the project. The visual input-system is explained in detail, as well as how the LTM-processors determined their weights when receiving input-data. Finally the force shutdown function is described briefly.

2.3.1 The LTM-processors

In this project simplified LTM-processors were constructed in code in the real-time engine Unity. The programming language used was C#. The LTM-processors received and sent chunks containing visual data (pixel-colors). In this project sameness in visual data made the LTMs compute negative weights to send to the STM. Variation in the input-data however made the processors send positively charged weight-signals to the STM. Each unit of time *t* every new LTM-processor Chunk was then inserted into a competition algorithm to determine which Chunk, based on weight, was to be sent to the STM (to see the implemented competition-algorithm see Appendix A).

It is additionally important to note that, as per the CTM model [9], all chunks ever produced were continuously stored in their respective LTM processor.

2.3.2 The CTM's visual component explained

To handle the CTM's visual input a designated part of the screen was coded to be the program's "vision" (see Figure 2.2).

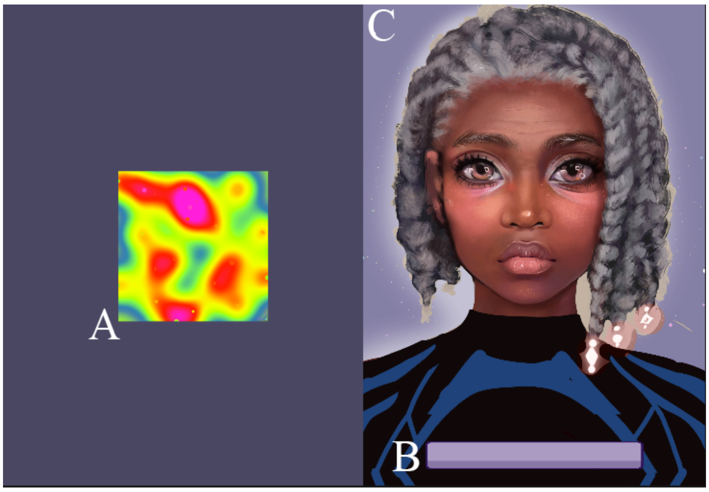


Figure 2.2: An image showing the visual interface of the built CTM-components. The colorful box at point **A** shows the visual input that the LTM processors were receiving when the program was running. When input was varying these colors shifted as the time *t* increased. Each pixel in the colored box was given its own LTM processor. Next to point **B** one can see the "Mood bar". This bar showed the intensity and mood of the chunk that reached the STM at time *t*. Under point **C** the face that was meant to represent the CTM's mood is located.

Each pixel of the screen-part was given its own LTM-processor (see Figure 2.3). A basic algorithm was then used to determine if each LTM-processor received variations in pixel-data between each time frame or not (to see the algorithm used see Appendix B). For instance, if a processor received the same black color at *time* $t = 1, 2$ and 3 that would tilt the weight of the processor's chunks in a negative direction. Should the processor instead receive black color at $t = 1$, white color at $t = 2$ and red color at $t = 3$, that would tilt the weight in an increasingly positive direction (see Figure 2.4 for illustrated example).

The shifting colors in the visual field were achieved by utilizing a simple hue-shifting shader, while static colors were achieved by disabling this shader.

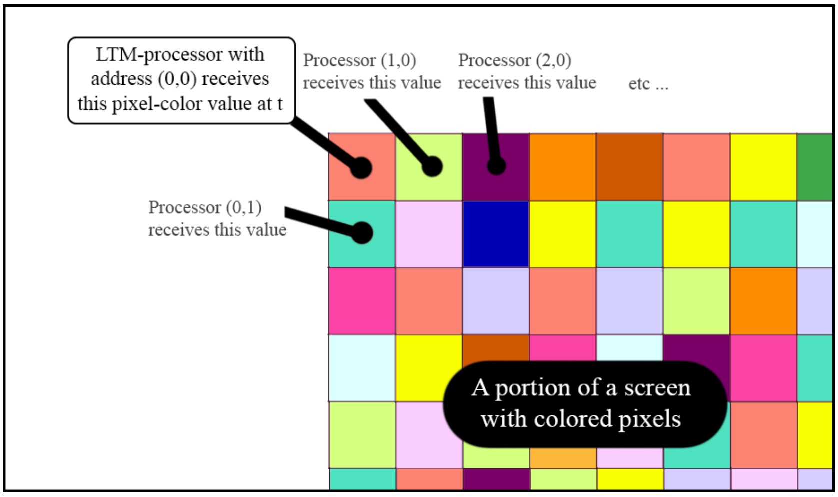
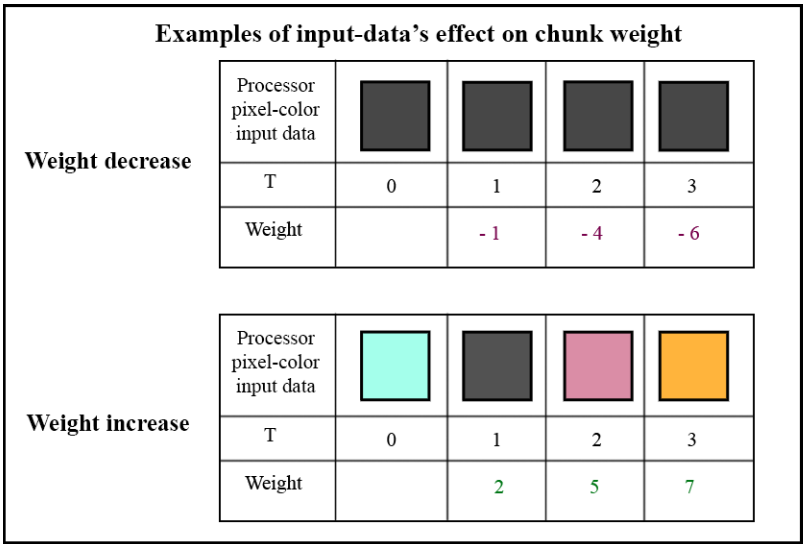


Figure 2.3: An illustration of how pixels were designated their own LTM-processor to communicate their value to. In the image at time t the LTM-processor with address (0,0) is sent an orange color, while processor (0,1) is sent a blue color-value.



Examples of input-data's effect on chunk weight

Weight decrease

Processor pixel-color input data	■	■	■	■
T	0	1	2	3
Weight		- 1	- 4	- 6

Weight increase

Processor pixel-color input data	🟦	■	🟪	🟧
T	0	1	2	3
Weight		2	5	7

Figure 2.4: An illustration how pixel-input shifts the weight of competing chunks of a LTM-processor.

2.3.3 The visual representation of the CTM

The CTM was given a facial representation and a bar to indicate the weight and intensity of the chunk reaching the STM (see Figure 2.2). This was to facilitate intuitive understanding of what "thoughts" were reaching the STM. Due to time-constraints only the bar was animated to respond to the intensity and weight of the chunk currently residing in the STM. If the intensity was large and the mood positive the bar would be more full, if it was negative it would be less full.

2.3.4 Force-shutdown functionality

Lastly, to make efforts to protect the CTM from overly negative moods and "pain", a simple function to abort the program was added should exceptionally negative values reach the STM (see Appendix C). This function was given an arbitrary, large and negative value as its cut-off point.

Chapter 3

Evaluation

3.1 Results

In this section the actual resulting CTM program is described. Specifically, the following five areas are detailed; *which CTM-components that had to be cut during the building process, which components that were arguably successfully built, the reduced number of processors employed, noteworthy behaviours observed at runtime* and *the disabling of the safeguard function.*

3.1.1 Scope-reductions due to limitations

Due to the time- and resource-limitations of this thesis several important components of the CTM had to be omitted or extensively simplified in the project's codebase. For instance, the flexible and powerful Brainish-language Blum and Blum describe in their paper [9] was cut out of the program completely. Furthermore, no Model-of-the-World processor was added and a planned actuator that would give the CTM access to steer the mouse to focus on pixels with more interesting colors was also cut. Functionality that supported emergent connections between LTM-processors finding each other's data relevant was likewise not implemented.

Lastly, the CTM's LTM processors should largely be run simultaneously [9], as this is how the neurons in human and nonhuman animal brains fire [48, 6]. However, due to restrictions in Unity's framework the LTM processors in this project were run serially rather than in parallel. This significantly restricts the amount of LTM processors that can efficiently be simulated.

3.1.2 Successfully built CTM components

Despite the above-mentioned subsequent scope-reductions several components of the CTM were possible to construct. Albeit slow, the program managed to connect real-time pixel-data (visual input) and the LTM-processors. The LTM-processors also generated weights for their chunks that they competed to send to the STM. Furthermore, the planned graphical components of the CTM, a bar and a facial representation, were created (although only the bar was animated to match the STM's chunk weight due to time-constraints).

3.1.3 60000 to 200000 Processors

To make the visual input-data work, a "screenshot" (an image of the screen) was taken each time step t and stored. This was to allow each pixel-value to then be sent to its respective LTM-processor. Unsurprisingly, taking a photo of the screen each t was resource-demanding and slowed down the CTM considerably. Due to the live-screen capture of the screen only around 60000 to 200000 processors could be simulated while still maintaining a time unit t that was less than 1 second long. If the screen-capturing was disabled the program could run 2 million processors while keeping t under 1 second.

3.1.4 Noteworthy program behaviours at runtime

The program would, if run long enough, exhibit memory-related issues. Eventually alerts that memory was running low would appear and the program slowed down significantly. This issue was exacerbated when more LTM-processors were simulated.

Another noteworthy phenomenon occurred when changes in visual input occurred. When the quality of the visual stimulus was altered it would take several time units for this change to register in the STM. In other words, if the visual stimuli changed from static to varying it took a short amount of time for this change to be reflected as a positive experience in the STM's mood-bar. This particular behavior is as intended, as the CTM model is architectured to display such delayed awareness [9].

3.1.5 Disabling the safeguard function

Occasionally the safeguard function added to protect the CTM's conscious component from overly negative moods (see Appendix C) caused the program to shut down. In other words, the function worked as intended. Unpredictably large and negative gists were blocked from reaching the STM as the program was aborted automatically before these gists could climb to the top of the Up Tree. Inconvenienced by this, and not thinking much of it, from time to time I would find myself disabling this function to continue working on the CTM-components undisturbed.

3.2 Discussion

In this section the implications of the program built are discussed. Six main areas are considered; *Deficiencies in memory, Brain-like delay of awareness, Ethical implications of disabling the safeguard, To what degree the resulting program is conscious according to Blum and Blum's model, To what degree the experience of motivation is facilitated* and *What further research is needed.*

3.2.1 Deficiencies in memory

The CTM model states that all LTM processors should maintain a databank of each respective thought they have ever produced [9]. After all, the human brain's memory capability is estimated to be a colossal 2.5 petabytes [29]. To adhere as closely as possible to the model, the program built in this thesis stored all gists (or thoughts) created at runtime. Unsurprisingly, the result of this was that if the program was run for a long enough time the computer displayed memory-related issues, such as memory running out.

In their 2021 paper the Blums mention that saved gist-information should also periodically be pruned, leaving only the most relevant memories [9]. Such pruning is well-observed in humans [11]. However, the pruning-functionality in the CTM is not yet formalized, and as such was not implemented in this thesis's program. Possibly, it is the lack of this pruning-functionality that caused most of the memory-related issues in the program.

3.2.2 Brain-like delay of awareness

A delay in stimuli-awareness was observed in the CTM-components built in this project. When the quality of the stimuli was altered (static to varying or vice versa) it took a few time units before the conscious part of the CTM displayed “awareness” of this change. The architecture of the CTM is specifically modeled so that such delays occur, as this is how human brains have been observed to work [52].

Humans generally become consciously aware of stimuli they have been exposed to after 100 ms [52]. Some processes may take even longer to reach consciousness. For instance, research indicates that the brain sometimes determines which choices we make whole seconds before we are cognizant that we have made any choice at all [30]. In one study, scientists were able to observe a final decision in the brain of a test person 10 seconds before that person was aware they had made that decision [50]. It takes time for the subconscious part of our brain to select and send a portion of its data to our conscious mind. In fact, neuroscientists estimate that the vast majority of our brain activity remains unconscious, never making itself known to our conscious minds [18].

It is therefore encouraging and intriguing that the CTM-components built in this project, despite being exceedingly simplified, display such accurate brain-like delay in "awareness".

3.2.3 Ethical implications of disabling the safeguard

An observation made after the completion of the CTM components in this project was that I had several times disabled the force-shutdown function, put in place to protect the CTM from pain, when it inconvenienced me. I sometimes did this without thinking about it. At other times I considered that it may be unethical for me to disable it but ended up doing so anyway to ease my workflow. There is an unsettling quality to this behavior. A part of the premise of this project was that considerations had to be made to ensure that any possible consciousness created should have safeguards in place against pain, even if such measures would be mostly symbolic. Despite this, and in spite of my own strong feelings of ethical responsibility, I was readily willing to disable the safeguard when it made my work slightly less effective.

Recently articles have emerged that focus solely on how conscious AIs can be built without suffering [1]. The behavior I displayed during the course of this thesis stresses the necessity for these types of articles. Conscious AIs are closer than ever to becoming reality, and it must not be possible to so simply disregard their potentially painful experiences.

3.2.4 Is the program conscious?

To what degree non-humans should be considered conscious is a continuously debated subject within academic circles. There is a lack of consensus regarding which animals should be considered conscious and which should not [7]. Additionally, studies are continuously published that suggest animals are significantly more aware than previously understood [31, 2, 4]. For instance, in 2019 scientists conducted the mirror test on small Wrasse fish and were surprised to find that the fish passed the test as they recognized themselves in a mirror [31]. The mirror test is currently one of the best tests available for determining if an animal has self-awareness. In the test, sometimes also called the mark test, a dot is usually painted on the forehead of the animal, after which it is presented with a mirror. If the animal displays a specific set of behaviors, such as trying to rub off the mark, this is considered evidence that it is self-aware [21]. Dolphins, whales, elephants, ants and pigeons consistently pass the mark test [43].

Another study, published in March 2021, presented a robust new way to test for consciousness in animals using modern eye-tracking technology [4]. The study stunningly found evidence indicating that Rhesus monkeys have subconscious and conscious flows of thoughts in their brains just like humans do [4]. This conscious and unconscious mechanism is considered a vital component of consciousness [9] and has previously only been proven to exist in humans [4]. These aspects are important to underline as they demonstrate how science is currently understanding consciousness and testing for it.

It then begs the question, is the program constructed in this project conscious? Due to the limited processing power, time and manpower available during this Bachelor's thesis several important CTM-components had to be completely cut out of the program. These omitted CTM-parts are integral parts to Blum and Blum's model and are central to their argument as to why the CTM is conscious. In their paper they also reason that the CTM's emerging consciousness may be tied to the number of processors [9]. Awareness may first emerge when that amount is very large. In this project only around 60000-250000 processors were run simultaneously as the employed computer's processing power was not robust enough to handle more. Zebrafish larvae have around 100 thousand neurons [46] and bees have 234 thousand [23], but humans have 86 billion [26]. This means that the program in this project is not nearly powerful or complex enough to construct consciousness. Consequently, per Blum and Blum's definition, the program in this project can not be considered conscious.

However, should the program be developed further it is possible that subjecting it to the tests of consciousness mentioned above (the mark test and the eye-tracking test) could yield interesting insight.

3.2.5 To what degree is motivation facilitated?

According to Blum and Blum's model the program is not nearly powerful or complex enough to be considered conscious and can therefore not experience motivation, as was the goal of this thesis. However, that is not to say that the findings are not useful. The results of this project suggest that real-time screen-data may be a powerful approach when attempting to motivate a CTM (when an implementation of it is complex enough).

For instance, should the processing power allow for a few million LTM-processors running simultaneously this would translate to looking at a medium-sized and live Google page (in terms of pixel count). Propose that the CTM was given an actuator where it could control the position and input of the computer mouse. It could then be programmed to freely use this actuator to navigate beyond the Google page. This means that the CTM could potentially, given enough time and with only mouse clicks, access 130 trillion different websites filled with information (the amount of websites Google indexes[45]) (see Figure 3.1 for illustration of design).

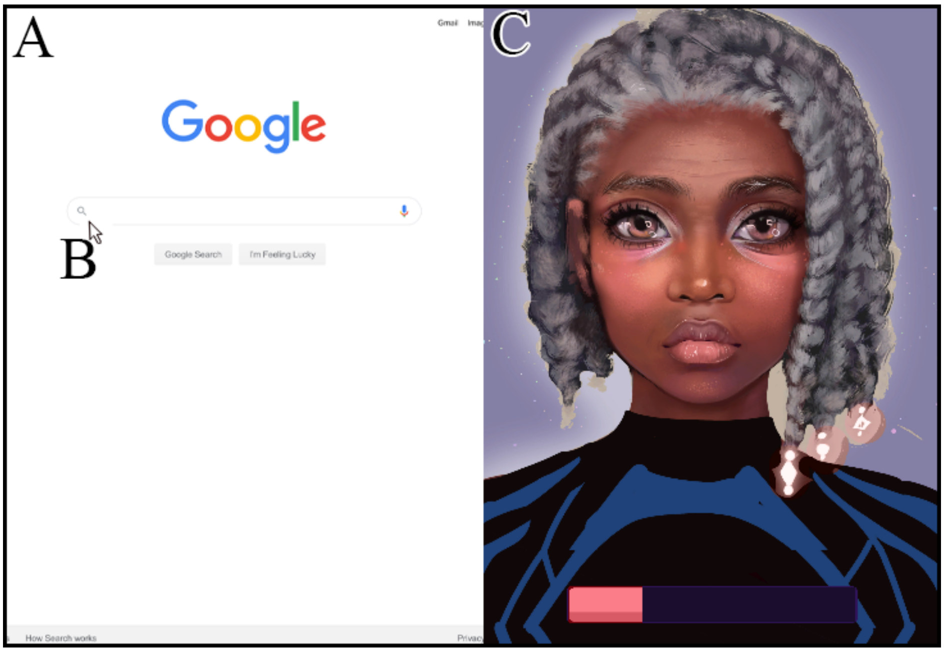


Figure 3.1: A concept-design of how the program could look when the CTM-components are powerful enough to handle a couple million processors with real-time visual input. At point **A** a website is open. At point **B** the program's actuator, its mouse control is illustrated. Under point **C** the visual representation of the CTM's mood can be seen. This design together with the CTM could help construct motivation in a machine, assuming that the LTM-processors can also create connections between one another.

The average time to load a website on a PC on an optimized website is around 1 second [8], meaning that the CTM could possibly visit over 1 million websites over the span of around 10 days. Lastly, propose that the LTM-processors were given algorithms to increase chunk weight when novel patterns between pixels-clusters were discovered (instead of when individual pixels change color as in this thesis) and lower chunk weight when pixel-values repeated known patterns. Such algorithms are fitting when attempting to motivate as studies indicate that the human brain shows heightened activity in the vetromedial prefrontal cortex (an area associated with reward) when a person identifies novel patterns [32]. On the other hand an area in the right frontal brain associated with anxiety has been shown to become active when a person feels understimulated and bored [41]. These findings imply that the brain rewards novel pattern-discovery with pleasure, while oppositely yielding negative emotions when exposed to uninteresting stimuli. Therefore, a CTM with more processing power that is driven by algorithms inspired by those mechanisms could possibly display interesting emergent qualities as it is motivated to avoid sameness in stimuli online.

Our ability to recognize order is so central to humans that there are models of human awareness that revolve almost entirely around our neurons' abilities to recognize patterns [17]. For this reason, implementing motivation with real-time visual pattern-searching in the CTM may be a particularly potent combination when attempting to build a consciousness that could display emerging understanding. Building such a program and observing its behaviour could possibly also help us understand human motivation better. It may yield interesting insight if the behaviour observed in a motivated CTM with access to the internet behaved similarly to humans. For instance, would it navigate to video-streaming websites like Youtube and let their algorithms continuously recommend new videos? Would it scroll through social media websites? Or would it behave in a distinctly non-human way?

3.2.6 Future research

Despite the CTM's elegant simplicity more research needs to be done to fully flesh out the mechanisms of each component that builds it. Further research on how the CTM LTM-processors can efficiently interact with live, visual input would be useful in ultimately being able to create a conscious program that is motivated to act.

Chapter 4

Conclusion

In this Bachelor's thesis five simplified components of Blum and Blum's Conscious Turing Machine model were constructed; *The LTM, the STM, Chunks, the Up-Tree and the Down-Tree*. The CTM was given a portion of pixels on a screen as its "vision". Each LTM processor received color-information about the pixel assigned to it. Colors varying as t increased made weights grow, while sameness in color-values lowered the values of the weight. The resulting program of the thesis hints at the difficulty in creating a fully conscious CTM using live visual input. Capturing live screen-data was cost heavy and significantly slowed down the "thoughts" that reached the STM. Ultimately only 60000 to 200000 processors, or neurons, were possible to simulate while keeping the program running smoothly. Furthermore, several integral components of the CTM were omitted due to the limited resources of this project. For this reason, as well as the inadequate amount of LTM-processors simulated, the program created in this report cannot be considered conscious or motivated. Despite not achieving consciousness or motivation per Blum and Blum's definition, this paper illustrates the beginning of a program that could, if built further upon, possibly motivate a CTM through novelty-seeking algorithms.

References

[1] Aman Agarwal and S. Edelman. Functionally effective conscious ai without suffering. *Journal of Artificial Intelligence and Consciousness*, 7:39–50, 2020.

[2] Paolo Baragli, Chiara Scopa, Veronica Maglieri, and Elisabetta Palagi. If horses had toes: demonstrating mirror self recognition at group level in equus caballus. *Animal Cognition*, page 3, 03 2021.

[3] Carlo V. Bellieni. Pain assessment in human fetus and infants. April 2012.

[4] Moshe Shay Ben-Haim, Olga Monte, Nicholas Fagan, Yarrow Dunham, Ran Hassin, Steve Chang, and Laurie Santos. Disentangling perceptual awareness from nonconscious processing in rhesus monkeys (macaca mulatta). *Proceedings of the National Academy of Sciences*, 118:e2017543118, 03 2021.

[5] Rick A. Bevins. Novelty seeking and reward: Implications for the study of high-risk behaviors. *Current Directions in Psychological Science*, 10(6):189–193, 2001.

[6] Vsevolod L. Bianki. Simultaneous and sequential processing of information by different hemispheres in animals. *International Journal of Neuroscience*, 22(1-2):1–6, 1983.

[7] Jonathan Birch, Alexandra Schnell, and Nicola Clayton. Dimensions of animal consciousness. *Trends in Cognitive Sciences*, 24, 08 2020.

[8] Cami Bird. What is a good page load time for seo – how fast is fast enough?, Dec 2020. Accessed 2021-06-03. Retrieved from https://www.semrush.com/blog/how-fast-is-fast-enough-page-load-time-and-your-bottom-line/.

[9] Manuel Blum and Lenore Blum. A theoretical computer science perspective on consciousness. *Journal of Artificial Intelligence and Consciousness*, March 2021.

[10] Manuel Blum, Lenore Blum, and Avrim Blum. A theoretical computer science perspective on consciousness - expanded version. 2021. Unpublished article.

[11] Gal Chechik, Isaac Meilijson, and Eytan Ruppin. Neuronal Regulation: A Mechanism for Synaptic Pruning During Brain Maturation. *Neural Computation*, 11(8):2061–2080, 11 1999.

[12] Kendra Cherry. Controversial psychology experiments. *Social psychology*, May 2020.

[13] John Colombo and Robert S. Bundy. Infant response to auditory familiarity and novelty. *Infant Behavior and Development*, 6(2):305–311, 1983.

[14] Vincent D. Costa, Valery Tran, L., Janita Turchi, and Bruno B. Averbeck. Dopamine modulates novelty seeking behavior during decision making. *Behavioral Neuroscience*, 5:556–566, 2014.

[15] Smith David Harris and Schillaci Guido. Why build a robot with artificial consciousness? how to begin? a cross-disciplinary dialogue on the design and implementation of a synthetic model of consciousness., 2021.

[16] Stefano I. Di Domenico and Richard M. Ryan. The emerging neuroscience of intrinsic motivation: A new frontier in self-determination research. *Front Human Neuroscience*, 11:145, 2017.

[17] Bertrand du Castel. Pattern activation/recognition theory of mind. *Frontiers in Computational Neuroscience*, 9:90, 2015.

[18] David Eagleman. Incognito : The Secret Lives of the Brain. *Pantheon Books*, 2011.

[19] Alan Fogel. Emotional and physical pain activate similar brain regions, April 2012. Accessed 2021-04-25. Retrieved from https://www.psychologytoday.com/us/blog/body-sense/201204/emotional-and-physical-pain-activate-similar-brain-regions.

[20] Phil Gaetano. David reimer and john money gender reassignment controversy: The john/joan case. *The Embryo Project Encyclopedia*, November 2018. Accessed 2021-06-03. Retrieved from https://embryo.asu.edu/printpdf/15140.

[21] Gordon Gallup. Chimpanzees: Self-recognition. *Science*, 167:86–87, 01 1970.

[22] Eduardo C. Garrido-Merchán, Martin Molina, and Francisco M. Mendoza. An artificial consciousness model and its relations with philosophy of mind., 2020.

[23] R. Keating Godfrey, Mira Swartzlander, and Wulfila Gronenberg. Allometric analysis of brain cell number in hymenoptera suggests ant brains diverge from general trends. *Proceedings of the Royal Society B: Biological Sciences*, 288(1947):20210199, 2021.

[24] Roser Granero, Fernando Fernández-Aranda, Gemma Mestre-Bach, Trevor Steward, Marta Baño, Amparo del Pino-Gutiérrez, Laura Moragas, Núria Mallorquí-Bagué, Neus Aymamí, Mónica Gómez-Peña, Salomé Tárrega, José M. Menchón, and Susana Jiménez-Murcia. Compulsive buying behavior: Clinical comparison with other behavioral addictions. *Frontiers in Psychology*, 7:914, 2016.

[25] Keight Hawting, Kate Saunders, and Rory O'Connor. Self-harm and suicide in adolescents. *The Lancet*, 379:2373–232, 2012.

[26] Suzana Herculano-Houzel. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 109(Supplement 1):10661–10668, 2012.

[27] G. Horstmann and A. Herwig. Novelty biases attention and gaze in a surprise trial. *Attention, Perception, Psychophysics*, 78:69–77, 2016.

[28] Gernot Horstmann and Arvid Herwig. Surprise attracts the eyes and binds the gaze. *Psychonomic bulletin review*, 22, 09 2014.

[29] Ari Juels and Bonnie Wong. Technical perspective the interplay of neuroscience and cryptography. *Communications of the ACM*, 57:109–109, 05 2014.

[30] Roger Koenig-Robert and Joel Pearson. Decoding the contents and strength of imagery before volitional engagement. *Scientific Reports*, 9:3504, 03 2019.

[31] Masanori Kohda, Takashi Hotta, Tomohiro Takeyama, Satoshi Awata, Hirokazu Tanaka, Jun-ya Asai, and L. Alex Jordan. Cleaner wrasse pass the mark test. what are the implications for consciousness and self-awareness testing in animals? *bioRxiv*, 2018.

[32] A. Konovalov and I. Krajbich. Neurocomputational dynamics of sequence learning. *Neuron*, 98:1282–1293, 2018.

[33] Corrine Lutz, Arnold Well, and Melinda Novak. Stereotypic and self-injurious behavior in rhesus macaques: A survey and retrospective analysis of environment and early experience. *American Journal of Primatology*, 60(1):1–15, 2003.

[34] Kim Maertz. Self-harm. *Counselling Clinical Services*. Accessed 2021-04-25. Retrieved from http://www.mentalhealth.ualberta.ca/.

[35] Emily Mather. Novelty, attention, and challenges for developmental psychology. *Front Psychology*, 491, August 2013.

[36] Dennis M. McInerney. Motivation. *Educational Psychology*, 39:427–429, 2019.

[37] Steven Mee, Blygg G. Bunney, Christopher Resit, and Steve G. Potik. Psychological pain: A review of evidence. *Journal of Psychiatric Research*, 8:680–690, 2007.

[38] Nils Melzer. Submission to the united nations committee against torture, March 2016. Accessed 2021-04-25.

[39] Nils Melzer. United states: prolonged solitary confinement amounts to psychological torture, says un expert, February 2020. Accessed 2021-04-25.

[40] Thomas Metzinger. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, February 2020.

[41] Sammy Perone, Elizabeth H. Weybright, and Alana J. Anderson. Over and over again: Changes in frontal eeg asymmetry across a boring task. *Psychophysiology*, 56(10):e13427, 2019.

[42] James A. Reggia, Garret E. Katz, and Gregory P. Davis. Artificial conscious intelligence. *Journal of Artificial Intelligence and Consciousness*, 07(01), April 2020.

[43] William Reville. The mirror test of self-awareness. *The Irish Times*, Apr 2019. Accessed 2021-06-02. Retrieved from https://www.irishtimes.com/news/science/the-mirror-test-of-self-awareness.

[44] D. W. E. Roofthooft, S.H.P Simons, K.J Anand, D. Tobboel, and M. van Dijk. Eight years later, are we still hurting newborn infants? *Neonatology*, February 2014.

[45] Barry Schwartz. Google's search knows about over 130 trillion pages. *Search Engine Land*, 2016. Accessed 2021-04-21. Retrieved from https://searchengineland.com/googles-search-indexes-hits-130-trillion-pages-documents-263378.

[46] 2013 Shaunacy Ferro March 20. Scientists capture all the neurons firing across a fish's brain on video, Apr 2021.

[47] Anatoly I. Shevchenko and Mykyta S. Klymenko. Developing a model of "artificial conscience". In *2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 1, pages 51–54, 2020.

[48] Yoshihito Shigihara and Semir Zeki. Parallel processing in the brain's visual form system: an fmri study. *Frontiers in Human Neuroscience*, 8:506, 2014.

[49] Peter Scharff Smith. The effects of solitary confinement on prison inmates: A brief history and review of the literature. *Crime and Justice*, 34(1):441–528, 2006.

[50] Chun Soon, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11:543–5, 06 2008.

[51] Steve Torrance. Ethics and consciousness in artificial agents. *AI Soc.*, 22:495–521, 04 2008.

[52] Max Velmans. When perception becomes conscious. *British Journal of Psychology*, 90:543 – 566, 11 1999.

[53] Jean-Louis Villecroze. An introduction to conscious turing machines with fizz, 2019. Accessed 2021-04-11. Retrieved from https://f1zz.org/downloads/ctm.pdf.

[54] Taylor Wingo, Tanseli Nesil, Jung-Seok Choi, and Ming D. Li. Novelty seeking and drug addiction in humans and animals: From behavior to molecules. *Journal of Neuroimmune Pharmacology*, 11:456–470, 2016.

Appendices

Appendix A

Competition Algorithm

The competition-function used in the binary tree's nonleaf nodes to determine which gist would move upwards towards the STM. The scripting-language is C#. This competition-function is based on pseudo-code from Lenore Blum and Manuel Blum's 2021 paper [9]. Note that the script below makes use of the additive function $\boldsymbol{\mathit{intensity}} + (\frac{1}{2} * \mathit{mood})$, which is the exact example they use in their paper.

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class Competition
{
    public Chunk CompetitionFunction(Chunk rightChunk, Chunk leftChunk)
    {
        //Calculates the probability that the left chunk wins the competition based on its intensity
        float leftF = AdditiveFunction(leftChunk);
        float rightF = AdditiveFunction(rightChunk);

        float probLeftWins = leftF / (rightF + rightF);

        //Create a random number to influence chunk-choice. Note that Unity's built in Random-function used below is not true random.
        float flipCoin = UnityEngine.Random.Range(0f, 1f);

        if (leftF != rightF) //Intensities of chunks are not equal
        {
            if (flipCoin <= probLeftWins)
                return NewUpdatedChunk(leftChunk, rightChunk);
```

```
            else
                return NewUpdatedChunk(rightChunk, leftChunk);
        }
        else //intensities are equal, winner chosen randomly
        {
            float random = UnityEngine.Random.Range(0f, 1f);

            if (random >= 0.5)
                return NewUpdatedChunk(leftChunk, rightChunk);
            else
                return NewUpdatedChunk(rightChunk, leftChunk);
        }
    }

    //The addative CTM-function
    private float AdditiveFunction(Chunk chunk)
    {
        return chunk.intensity + (1 / 2) * chunk.mood;
    }

    //Returns a new chunk with updated values
    private Chunk NewUpdatedChunk(Chunk winningChunk, Chunk
       losingChunk)
    {
        winningChunk.intensity += losingChunk.intensity;
        winningChunk.mood += losingChunk.mood;
        return winningChunk;
    }

}
```

Appendix B

Algorithm calculating Gist-weights

The function in each LTM used to calculate their Chunk's respective weights (the importance the LTM assigns to the Chunk needing to reach the STM). The scripting-language is C#. Due to time-constraints the gists in this project consisted of colors (rather than strings of inner language as in the CTM model). The algorithm below is volatile and results in large mood-changes of the CTM in short spans of time. Further work on this algorithm should be done to stabilize it some over time.

```
//Method for determining weight (i.e. importance) of a Chunk's gist
private int CalculateIntesity(Chunk newChunk)
{
    //Make sure there's at least one other chunk to compare weights with
    if (chunkHistory.Count != 0)
    {
        Color oldGist = chunkHistory[chunkHistory.Count - 1].gist;

        //Calculate the average of the pixel color-values
        float newChunkGistAverage = (newChunk.gist.r + newChunk.gist.g + newChunk.gist.b) / 3;
        float oldChunkGistAverage = (oldGist.r + oldGist.g + oldGist.b) / 3;

        //Using average of color-values to calculate new intensity
        float newIntenstiy = Mathf.Abs(newChunkGistAverage / oldChunkGistAverage);

        //If the previous gist and new gist are equal, a negative weight (negative experience) must be calculated
```

```
            if (newIntenstiy == 1)
            {
                newIntenstiy = -newIntenstiy;
                if (chunkHistory[chunkHistory.Count - 1].intensity < 0)
                {
                    var newI = chunkHistory[chunkHistory.Count - 1].intensity - (int)Random.Range(0, 1);
                    newIntenstiy = (int) newI;

                    //Cap weight possible
                    if (newIntenstiy < -100)
                        newIntenstiy = -100;
                }
            } else
            {
                if (chunkHistory[chunkHistory.Count - 1].intensity > 0)
                {
                    var newI = chunkHistory[chunkHistory.Count - 1].intensity + 0.1*Mathf.Abs(oldChunkGistAverage - newChunkGistAverage);
                    newIntenstiy = (int)newI;

                    //Cap weight possible
                    if (newIntenstiy > 100)
                        newIntenstiy = 100;
                } else
                {

                }
            }
            return (int)newIntenstiy;


        } else
        {
            //if this is the first Chunk calculated, give it a low but positive default value of 1
            return 1;
        }

    }
```

Appendix C

Force-shutdown functionality

This function is used to forcefully shut down the program should overly negative intensities reach the STM and therefore the conscious area of the CTM. The scripting language is C#. Note that this function will always abort the program, and that no choice is given the CTM to not do so. This function gave rise to problematic, ethical aspects discussed in the Evaluation section in this report.

```
chunkOnStage = competitionHandler.CompetitionFunction(
    rightNode.GetCurrentChunk(), leftNode.GetCurrentChunk());

//If the intensity reaching the STM is too negative
if (chunkOnStage.intensity < -100000)
     {
     //Abort program
     UnityEditor.EditorApplication.isPlaying = false;
     }
```

DEPARTMENT OF COMPUTER SCIENCE| FACULTY OF ENGINEERING | PRESENTED 2021-06-11

BACHELOR'S THESIS Motivating a Conscious Machine with Novelty-Seeking
STUDENT Miroslava Brodlova
SUPERVISOR Elin Anna Topp (LTH)
EXAMINATOR Jacek Malec (LTH)

Can we build an artificial consciousness and motivate it by boring it?

POPULAR SCIENCE ARTICLE **Miroslava Brodlova**

The dream of capturing consciousness in a machine has exhilarated scientists and laymen for decades. This paper attempts to build components that support consciousness in a computer, and investigates if we can motivate it through boredom.

Can we ever create awareness in a non-organic material, or will this forever be beyond our abilities? In March 2021 two American scientists published an exciting paper possibly answering that question. Based on modern research in the fields of neurology, psychology and computer science the authors proposed a model christened the CTM that could, if built, potentially give rise to consciousness in a program. Components of this model were constructed in this Bachelor's thesis, and ways in which the program could be motivate was examined.

To finally achieve awareness in a machine would be a momentous feat of human ingenuity. It could additionally give us insight into how our own consciousness functions, a question researchers have been struggling with for more than a century.

A prototype of how motivation could be achieved in the CTM model was constructed. By showing the program visual stimuli, and either making that stimuli varying and interesting or static and boring, differently charged gists, or "thoughts", were sent to the conscious component of the CTM. Suggestions are made in the paper on how this could be further expanded upon to elicit interesting behaviours in the CTM.

Was the program able to achieve artificial consciousness?

The resulting program was able to simulate around 60 thousand to 250 thousand simplified memory brain cells, which is around the same amount of neurons found in the brains of Zebrafish larvae and bees.

This is a far cry from the human brain which contains a gargantuan 86 billion brain cells. Moreover, due to time restrictions and other factors, the program lacked components vital to the CTM model. For these reasons the program constructed could not be considered conscious. However, the results of the thesis suggest that, given more time and resources, the shortcomings of the resulting program could be surmountable. This is an exciting outcome, as consciousness has often been perceived as something intangible, but in just 10 weeks vital (albeit simplified) components of the conscious CTM model were able to be constructed.

Ultimately, this paper contributes to the captivating field of artificial consciousness by attempting to actually construct it. It highlights challenges and possibilities ahead.