FOM Hochschule für Oekonomie & Management

Study Center Stuttgart

Bachelor-Thesis

in the study program Wirtschaftsinformatik - Business Information Systems

to obtain the degree of

Bachelor of Science (B.Sc.)

on the topic

Is it possible to give emotions to artificial intelligence?

by

Anna Lena Katharina Braun

Matriculation number:	494571
First assessor:	Hamid, Maher
Submission date:	2021-11-04

Inhaltsverzeichnis

Table of Figures IV

Table of Source Code Listings V

List of Abbreviations VI

1 Introduction 1

1.1 Problem 1

1.2 Goal 2

2 Theoretical background 4

2.1 Artificial Intelligence 4

2.1.1 Definition 4

2.1.2 Current state-of-the-art 4

2.1.3 Limitations 9

2.1.4 Used model 12

2.2 Emotions 15

2.2.1 Definition 15

2.2.2 Emotions in evolution 16

2.2.3 Anatomy of the Amygdala 17

2.3 Fear 19

2.3.1 Definition 19

2.3.2 Innate fear 20

2.3.3 Learned fear 26

3 Creation of the fear circuit code 36

3.1 Method 36

3.2 Mapping 38

3.3 Logical composition 43

3.4 Creation of the code 44

4 Implementation of Fear into an AI 61

4.1 Integration into an AI 61
4.2 Interplay of the fear circuit and the AI 64
5 Limitations 67
5.1 Remapping onto a computational environment 67
5.2 Neurological implications 71
6 Conclusion 73
Appendix 76
Bibliography 77

Table of Figures

Figure 1: Neural network for classifying images of lamas and giraffes 7
Figure 2: One Pixel Attack 10
Figure 3: Conversation with Telegram Bot on the Computer Console 13
Figure 4: Conversation with the Telegram Bot on Telegram 14
Figure 5: Anatomy of fear conditioning circuits in the brain 18
Figure 6: Schematic representation of the neural circuits mediating innate fear 22
Figure 7: Classical Conditioning Diagram 27
Figure 8: Overview of the brain areas reproduced by the model of learned fear 30
Figure 9: Learned fear model functioning during the tests 32
Figure 10: Conversation with the Telegram Bot on the Console part 1 64
Figure 11: Conversation with the Telegram Bot on the Console part 2 65
Figure 12: ISO/IEC 25010 Software Quality Pillars 69

Table of Source Code Listings

Listing 1: Command to create test_table in MySQL 45
Listing 2: Modifications needed when tested on other devices..............................45
Listing 3: Import declarations in fear_circuit.main() 46
Listing 4: Main function part 1 46
Listing 5: Main function part 2 47
Listing 6: The class UnconditionedStimuli..48
Listing 7: The class ConditionedStimuli..50
Listing 8: Main function part 3 52
Listing 9: Main function part 4 53
Listing 10: Functions for fear responses 55
Listing 11: Main function part 5 58
Listing 12: Main function part 6 59
Listing 13: Command to run Telegram Bot on Telegram.......................................61
Listing 14: Command to run Telegram Bot on the Console...................................61
Listing 15: Evaluation of the communication channel..62
Listing 16: Modification in gpt2bot.console_bot.run() 62
Listing 17: Command to install requirements 63

List of Abbreviations

ACL	Annual Meeting of the Association for Computational Linguistics
AH	Anterior Hypothalamic Nucleus
AI	Artificial Intelligence
ANIPS	Advances in Neural Information Processing Systems
ANYAS	Annals of the New York Academy of Sciences
AOS	Accessory Olfactory System
API	Application Programming Interface
ARN	Annual Review of Neuroscience
AuC	Auditory Cortex
BA	Basal Nucleus of the Amygdala
BLA	Basolateral Amygdala
BMA	Basomedial Amygdala
Brit J Psych	British Journal of Psychiatry
CEA	Central Nucleus of the Amygdala
CeL	Lateral Subdivision of the Central Amygdala
CeM	Medial Subdivision of the Central Amygdala
CoA	Cortical Amygdala
Cog Emot	Cognition and Emotion
CoRR	Computing Research Repository
CPU	Central Processing Unit
CS	Conditioned Stimulus
Dia C Neuro	Dialogues in Clinical Neuroscience
Em Rev	Emotion Review
Fr Sys Neuro	Frontiers in Systems Neuroscience
Fr Psych	Frontiers in Psychology

GIF	Graphics Interchange Format
GPT-3	Generative Pre-trained Transformer 3
IC	Inferior Colliculus
ICSE	International Conference on Software Engineering
IEEE TEC	IEEE Transactions on Evolutionary Computation
IJCNN	Proceedings of International Joint Conference on Neural Networks
IL	Infralimbic Cortex
ITC	Intercalated Cells of the Amygdala
J Intell	Journal of Intelligence
Jour E P	Journal of Experimental Psychology
Jour O T	Journal of Object Technology
LA	Lateral Nucleus of the Amygdala
Learn Mem	Learning & memory
MEAdd	Posterior Dorsal Portion of the Medial Amygdala
MEApv	Posterior Ventral Portion of the Medial Amygdala
MiMa	Minds and Machines
ML	Machine Learning
MND	Mind
MOS	Main Olfactory System
MPN	Medial Preoptic Nucleus
MTN	Midline Thalamic Nuclei
N NEURO	Nature Neuroscience
N R NEURO	Nature Reviews Neuroscience
Neb S M	Nebraska Symposium on Motivation
NLP	Natural Language Processing
NEU B REV	Neuroscience and biobehavioral reviews

NRN	Neuron
OHY	One Hundred Year Study on Artificial Intelligence (AI100)
PAG	Periaqueductal Gray
PAGd	Dorsal Periaqueductal Gray
PAGvl	Ventrolateral Periaqueductal Gray
PB	Parabrachial Nucleus
Phil TRS B	Philosophical Transactions of the Royal Society B
Phys Rev	Physiological Reviews
PL	Prelimbic Cortex
PMDdm	Dorsomedial Portion of the Dorsal Premammillary Nucleus
PMV	Ventral Premammillary Nucleus
Psych Tho	Psychological Thought
RAM	Random Access Memory
RGN	Retinal Ganglion Cells
SC	Superior Colliculus
SCI	Science
SQL	Structured Query Language
SYN	Synapse
US	Unconditioned Stimulus
VMHvl	Ventrolateral Portion of the Ventromedial Hypothalamic Nucleus
V1	Primary Visual Cortex
W IRC Sci	Wiley Interdisciplinary Reviews: Cognitive Science
Wo Str	Work and Stress

1 Introduction

1.1 Problem

“Emotions enrich; a model of mind that leaves them out is impoverished”.[1]

This quote stems from Daniel Goleman, author of the bestseller Emotional Intelligence, whose release introduced the importance of emotions for human intelligence.[2] Emotions are crucial for human survival.[3] Emotions are the reason why the eleven-year-old Andrea was saved by her parents during a train accident, while the train was sinking into a river.[4] Whereas her parents died, she could survive.[5] The love for their cherished daughter outweighed the parents’ instincts of survival.[6] The moment a human eye sees a snake, an emotional response is triggered in our brain even before we understand what is happening. [7] Our heart rate accelerates, blood pressure increases, and muscles prepare for action.[8]

Fear particularly manifests the importance of emotions over other areas of the brain. According to the neuroscientist Kay Tye, “fear resembles a dictator that makes all other brain processes (from cognition to breathing) its slave”.[9] The Amygdala, a brain structure and center of the fear circuit,[10] initiates emotional reactions and emotional memories, without any conscious or cognitive participation.[11]

In short: emotions impact our actions and perceptions of the world. They are vital to human survival[12] and human intelligence, which is the greatest known intelligence.[13] Thus, it can be assumed that any non-biological intellect, aiming to increase its own intelligence, should have emotional circuits like human’s.

This idea was mentioned by Alan Turing in his popular 1950 paper called Computing Machinery and Intelligence.[14] There, he includes an argument on how machines

[1] *Goleman, D*., Emotional Intelligence, 1996, p. 45.
[2] Cf. *Picard, R., W.,* Affective Computing, 2015, p.17.
[3] Cf. *LeDoux, J.,* Emotion, 2012, p. 660.
[4] Cf. *New York Times*, Emotional Behavior, 1993, no page found.
[5] Cf. ibid.
[6] Cf. *Goleman, D*., Emotional Intelligence, 1996, p. 4.
[7] Cf. *Goleman, D*., Emotional Intelligence, 1996, p. 22.
[8] Cf. ibid.
[9] *Mobbs, D. et al.,* Fear, 2019, p.1206.
[10] Cf. Armony, J.: Emotion, 2005, p. 1598.
[11] Cf. *Goleman, D*., Emotional Intelligence, 1996, p. 21.
[12] Cf. *LeDoux, J.,* Emotion, 2012, p. 660.
[13] Cf. *Erickson, R*., Intelligence, 2014, p. 119-120.
[14] Cf. *Turing, A.,* Computing, 1950, p.442.

would need to feel emotions to fully replicate the human brain.[15] State-of-the-art models in AI (Artificial Intelligence = AI) simulate biological neural networks[16] and use elements from statistics, like Bayesian statistics.[17] Hence, there are no high-level abstractions of the brain circuits involved in emotions implemented in artificial agents. Yet, the goal of AI development is to design machines that mimic human minds.[18]

1.2 Goal

Given the problem described in chapter 1.1, the following research question is established:

Is it possible to implement the emotion fear into Artificial Intelligence?

The goal of this thesis is to pave the way for research on how to implement further brain circuitry into artificial agents to make them more intelligent. Hence, it should establish a new branch in the creation of AI. To achieve this goal, it is analyzed whether emotions can be implemented into AI. This idea is specifically tested on the emotion fear and its underlying fear circuits. Both, the creation of a computational fear circuit and its implementation into a state-of-the-art AI model will determine the success of this research.

This topic is highly relevant given the multitude of highly respected businessmen, politicians and researchers valuing the development of Artificial Intelligence as the most important development for humanity. Sergey Brin, the co- founder of Google says Artificial Intelligence is the most important computational development in his lifetime.[19] Elon Musk, co-founder of SpaceX and Tesla,[20] says AI is the biggest threat to humanity.[21] Russia's president Vladimir Putin argues that whoever will become the leader in AI will rule the world.[22] The theoretical physicist Stephen Hawking has warned that developing a superintelligent AI will mean the end of humanity.[23]

---

[15] Cf. ibid.
[16] Cf. *Russell, S.,* Artificial Intelligence, 2020, p. 290.
[17] Cf. *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 307.
[18] Cf. *Minsky, M.,* Emotion, 2007, p.6.
[19] Cf. *Brin, S.,* Google, 2017, no page found.
[20] Cf. *Tesla,* Elon Musk, n.d., no page number.
[21] Cf. *Shead, S.,* Elon Musk, 2020, no page found.
[22] Cf. *RT.com*, Putin, 2017, no page found.
[23] Cf. *Cellan-Jones, R.,* Hawking, 2014, no page found.

As fear is a fundamental requirement of survival[24], answering whether it is possible for AI to feel fear is essential to further evaluate the consequences of developing AI.

To achieve this goal, characteristics and limitations of state-of-the-art AI models will be examined. The AI model, into which the fear circuit code will be implemented, is presented in chapter 2.1. A theoretical foundation of the term emotions will be provided in chapter 2.2, with an emphasis of the anatomy of the Amygdala, priming the reader for the deep, theoretical analysis of the two fear conditions, described in the following subchapter 2.3. Modern computational models will be used to understand the most important phenomena associated with fear. Then, the research method for creating the fear circuit code is presented in chapter 3.1, alongside a mapping of fear components from the biological onto the computational environment in the following subchapter. The theoretical insights about fear are combined with the results of the remapping in a logical composition of the fear circuit in chapter 3.3. This composition is transferred into program code in 3.4 and integrated into the chosen AI model in 4.1. Chapter 4.2 analyzes the symphony of both program codes and chapter 5 outlines the limitations of the approach. This ensures that researchers can clearly understand the benefits and challenges of the approach and develop it further. All key insights of this study are summarized in the conclusion.

[24] Cf. *LeDoux, J.*, Emotion, 2012, p. 676.

2 Theoretical background

2.1 Artificial Intelligence

2.1.1 Definition

Definitions of AI are very diverse. They range from “the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence”[25] to "machines with cognitive skills that rival or surpass those of humanity”.[26] Stanford’s One Hundred Year Study on Artificial Intelligence from 2016 describes AI as a field that inevitably loses claim of its acquisitions over time.[27] As new technologies emerge and people become accustomed to them, they are no longer considered as AI.[28] For example, a calculator would theoretically fit into the definition of AI. [29] Yet, as calculators have experienced a wide-spread adoption many years ago and more advanced tools like smartphones have emerged, they are not considered as AI nowadays.[30] Thus, instead of delivering one life-changing invention, AI technologies are incrementally improving over time, supplanting exiting technologies once considered as AI.[31]

The goal of AI research is to engineer a system that can learn and execute any given task on its own, from teaching molecular biology to school children to running a government.[32] It should be able to do so by acquiring knowledge from available resources, by asking questions and through developing and executing plans.

2.1.2 Current state-of-the-art

A big branch of AI is ML (Machine Learning = ML).[33] It is composed of computational methods used to make predictions or improve performance by experience.[34] The systems gain experience through analyzing data from the past.[35] Thus, ML is heavily related to data science and statistics.[36] The quality and size of these datasets are

[25] *High, P.,* Artificial Intelligence, 2017, no page found.
[26] *Armstrong, S.,* Artificial Intelligence, 2017, p. 1.
[27] Cf. Stone, P. et al., AI Report, 2016, p. 12.
[28] Cf. ibid.
[29] Cf. ibid.
[30] Cf. ibid.
[31] Cf. ibid.
[32] Cf. *Russell, S.,* Artificial Intelligence, 2020, p. 46.
[33] Cf. *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 3-4.
[34] Cf. *Mohri, M. et al.,* Machine Learning, 2018, p. 1.
[35] Cf. ibid.
[36] Cf. ibid.

fundamentally important to ensure systems leveraging ML make predominantly right predictions.[37]

Common learning tasks in ML scenarios are classification, ranking, clustering, and dimensionality reduction.[38] Classification means objects are assigned to categories.[39] For example, a picture of an animal can be classified as a cat. Regression aims at predicting the real value of an item.[40] For example, a ML model using regression can be used to predict stock prices. Ranking scenarios order items depending on some criterion.[41] These algorithms can be used to make search engine result pages or social media feeds.[42] In clustering, a set of items is partitioned into homogeneous subsets.[43] For example, a retailer might cluster its customers based on their profiles to enable targeted marketing campaigns.[44] Dimensionality reduction is a popular technique to preprocess images for computer vision tasks.[45] It transforms the initial representation of an item like pictures into a lower-dimensional representation.[46] Important properties of the initial item are preserved or enhanced.[47]

The most common learning scenarios to train a ML model are supervised learning, unsupervised learning, and reinforcement learning.[48] Supervised learning is often used for classification, ranking and regression.[49] ML models trained in a supervised learning setting are exposed to training examples which contain relevant information, in form of labels.[50] Then, it predicts the labels of unseen items.[51] For example, it could be trained on thousands pictures of cats and dogs, labeled accordingly.[52] Then, it will be able to classify new pictures it has not seen before as a cat or dog.[53]

---

[37] Cf. ibid.
[38] Cf. *Mohri, M. et al.,* Machine Learning, 2018, p. 3.
[39] Cf. ibid.
[40] Cf. ibid.
[41] Cf. ibid.
[42] Cf. ibid.
[43] Cf. ibid.
[44] Cf. *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 264.
[45] Cf. *Mohri, M. et al.,* Machine Learning, 2018, p. 3.
[46] Cf. ibid.
[47] Cf. ibid.
[48] Cf. *Mohri, M. et al.,* Machine Learning, 2018, p. 6.
[49] Cf. ibid.
[50] Cf. *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 4.
[51] Cf. *Mohri, M. et al.,* Machine Learning, 2018, p. 6.
[52] Cf. ibid.
[53] Cf. ibid.

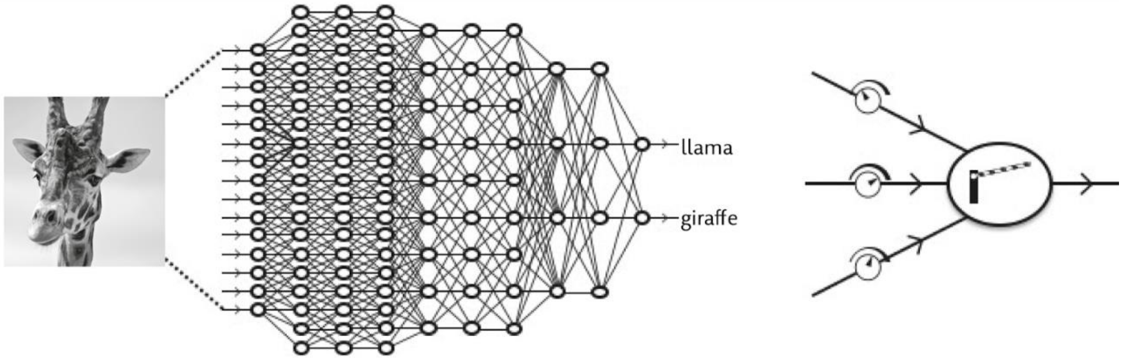
By contrast, the ML model learns patterns in the training data without any labels or feedback in unsupervised learning.[54] Thus, there is no distinction between labeled training data and unlabeled test data, as with supervised learning.[55] For example, an agent could be trained on animal pictures and create clusters for all similar images, like cats or dogs.[56] Unsupervised learning is used for clustering or dimensionality reduction.[57]

In Reinforcement learning, training and testing phases are mixed.[58] By interacting with its environment, the agent collects information and receives a reward or punishment for this action.[59] The agent's objective is to maximize its reward.[60] Hence, the agent faces a dilemma of choosing to explore to learn from reinforcements of its own actions, rather than plain data sets.[61] Reinforcement learning enables agents driven by their own experiences.[62]

An important algorithmic paradigm of ML are neural networks.[63] The term network refers to biological neural networks, which inspired the creation of artificial neural networks[64]. In simplified terms, the human brain consists of many interconnected neurons, which form a large communication network that enables the brain to perform highly complex computations.[65] Artificial neural networks are computational constructs, which model this paradigm.[66] They are complex mathematical expressions that consist of many layers and nodes.[67] Each node corresponds to an expression which is adjusted by changing the weights on each input, as shown on the left side of figure 1.[68]

[54] Cf. *Russell, S., Norvig, P.*, Artificial Intelligence, 2021, p. 826-827.
[55] *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 4.
[56] Cf. *Russell, S., Norvig, P.*, Artificial Intelligence, 2021, p. 826-827.
[57] Cf. *Mohri, M. et al.,* Machine Learning, 2018, p. 6.
[58] Cf. *Russell, S.., Norvig, P.,* Artificial Intelligence, 2021, p. 840.
[59] Cf. ibid.
[60] Cf. ibid.
[61] Cf. ibid.
[62] Cf. ibid.
[63] Cf. *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 228.
[64] Cf. *Russell, S.,* Artificial Intelligence, 2020, p. 288-290.
[65] Cf. *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 228.
[66] Cf. ibid.
[67] Cf. *Russell, S.,* Artificial Intelligence, 2020, p. 288-290.
[68] Cf. ibid.

Figure 1: Neural network for classifying images of lamas and giraffes



Source: *Russell, S.,* Artificial Intelligence, 2020, p. 289

The figure visually illustrates the mathematical expressions in a neural network that classifies pictures as llamas and giraffes.[69] The weight determines how much attention the node pays to an incoming value, to suppress small signals and allow large ones to go through.[70] Hence, the weight could be compared to volume controls, which filter out unnecessary noise.[71] On the left side of the network, image pixel values are fed as input into the network.[72] Two nodes serve as the output of the neural network on the right side, indicating the possibility of a given image to be a giraffe or llama[73]. Hence, the learning process consists of adjusting the weights in the network, leading to a reduction of prediction errors on labeled examples[74]. Artificial neural networks with more than two layers are called deep neural networks and their associated paradigm is called deep learning.[75] Currently, deep learning is the most popular technology used in AI.[76] The ImageNet competition in 2012 kick-started the wide popularity of deep learning, although the technology itself was invented long before.[77] In this competition, over one million pictures had to be classified into one of a thousand categories by an algorithm.[78]

[69] Cf. ibid.
[70] Cf. ibid.
[71] Cf. ibid.
[72] Cf. ibid.
[73] Cf. ibid.
[74] Cf. ibid.
[75] Cf. *Shalev-Shwartz, Shai, Ben-David, Shai,* Machine Learning, 2014, p. 240.
[76] Cf. *Russell, S.,* Artificial Intelligence, 2020, p. 290.
[77] Cf. Krizhevsky, A. et al., ImageNet, 2012, p. 1106.
[78] Cf. ibid.

There are different use cases that can be tackled through Machine Learning, like Computer Vision or NLP (Natural Language Processing = NLP).[79] Computer Vision is often used in combination with deep learning, as it was the case in the ImageNet competition[80]. While the lowest error rate achieved in this contest was 15,3%, state-of-the-art Computer Vision models have an accuracy of approximately 2%.[81] As the error rate of trained humans is around 5%, this demonstrates the reliability of these systems.[82] Its application ranges from self-driving cars to evaluating and grading cucumbers.[83] Like Computer Vision, NLP heavily relies on Deep Learning. NLP is often used in the context of speech recognition and machine translation.[84] For example, a machine translation application might translate an English sentence, which is its input, into a Japanese sentence.[85] Since 2020, state-of-the-art machine translation systems have a human-level performance for language pairs that offer many paired datasets, like English and French.[86] By training machine translation models on multiple languages, they can transfer the internal meanings from one language to another.[87] A model might be trained to translate Portuguese into English and later it will be trained to translate English into Spanish.[88] In these circumstances, the model can translate Portuguese to Spanish without being trained on English to Spanish language sets.[89] There are different types of architectures of the neural networks underlying NLP models, like recurrent, convolutional or transformer networks.[90] Transformer models were described by Google researchers for the first time in 2017.[91] This architecture results in a significantly better model accuracy and less training time compared to the other two architectures.[92] The transformer architecture relies on an attention mechanism drawing dependencies between the model's input and output.[93]

---

79 Cf. *Mohri, M. et al.,* Machine Learning, 2018, p. 2.
80 Cf. *Russell, S., Norvig, P.,* Artificial Intelligence, 2021, p. 44.
81 Cf. ibid.
82 Cf. ibid.
83 Cf. ibid.
84 Cf. ibid.
85 Cf. ibid.
86 Cf. ibid.
87 Cf. ibid.
88 Cf. ibid.
89 Cf. ibid.
90 Cf. *Vaswani, A.,* Transformer, 2017, p. 5999.
91 Cf. ibid.
92 Cf. *Vaswani, A.,* Transformer, 2017, p. 6002.
93 Cf. *Vaswani, A.,* Transformer, 2017, p. 5999.

2.1.3 Limitations

Gary Kasparow, long-time world champion in chess, was beaten by IBM's computer DeepBlue in 1997.[94] Of course, this has not been the first-time which humans were defeated by machines.[95] Cars are faster than the best human sprinters and a calculator can calculate faster than any mathematician.[96] However, this moment was historic as chess is considered as a highly intellectual pursuit and intellect itself is viewed as the ultimate differentiation between humans and other species.[97]

Yet, AI is nowhere near human intelligence[98]. Artificial agents might be superior at very specific tasks, but they humans are generalists that can execute and learn a variety of very different tasks.[99] A human chess player can cook dinner, drive a car and write books, whereas a narrow AI cannot transfer its skills to different tasks.[100] A state-of-the-art AI designed for playing chess is not capable of writing an essay about chess or fundamental tasks like applying for a graduate program.[101]

Although AI systems, as in the case of DeepBlue, outperform humans in specific tasks, they are susceptible to errors rare in humas. In 2019, researchers demonstrated that by altering one pixel in natural images, deep neural networks are likely to misclassify the entire image.[102] Some of their findings are shown in figure 2. The single pixels are encircled in red.[103] The black words show the original classification and confidence of the algorithm for a specific label and the blue words show its classification and confidence level after the one-pixel attack.[104] By contrast, humans would not have any problems recognizing a teacup as a teacup after one pixel of the image has been altered.

[94] Cf. *Kurzweil, R.,* Artificial Intelligence, 2006, p. 2.
[95] Cf. ibid.
[96] Cf. ibid.
[97] Cf. ibid.
[98] Cf. *Kurzweil, R.,* Artificial Intelligence, 2006, p. 3.
[99] Cf. ibid.
[100] Cf. ibid.
[101] Cf. ibid.
[102] Cf. *Su, J. et al.,* One Pixel, 2019, p. 828-829.
[103] Cf. ibid.
[104] Cf. ibid.

Figure 2: One Pixel Attack

Source: *Su, J. et al.,* One Pixel, 2019, p. 829

Another problem of current AI systems is their high demand in resources. For example, training GPT-3 (Generative Pre-trained Transformer 3 = GPT-3) cost around $12 million.[105] This autoregressive language model produces human-like text through deep learning techniques.[106] In an official blog post by Microsoft, who has an exclusive license for GPT-3, it is described as “the largest and most advanced language model in the world”[107]. Yet, it needs 175 billion parameters[108] and millions of money to be able to write code sequences or human-like Wikipedia entries[109]. By contrast, humans need a fraction of these resources to acquire the same skillset, which shows that the current deep learning approach might lead to great results, but its efficiency is expandable.

[105] Cf. *Floridi, L., Chiriatti, M.,* GPT-3, 2020, p. 684.
[106] Cf. ibid.
[107] *Scott, K.,* GPT-3, 2020, no page number.
[108] Cf. ibid.
[109] Cf. *Floridi, L., Chiriatti, M.,* GPT-3, 2020, p. 684.

The leading AI researcher Stuart Russell argues that the current state-of-the-art AI systems are nowhere close to reaching human-level intelligence.[110] Although the human brain and deep neural networks are both made up of neurons that are arranged in circuits, collecting more neurons will not lead to more intelligence.[111] Instead, the underlying circuits must be arranged in certain ways and additional software layers must be added.[112]

Demis Hassabis, the CEO of Google DeepMind, a company focused on developing human-level intelligence, argues that deep learning is "definitely not enough to solve AI".[113] From his point of view, deep learning is equivalent to the sensory cortices in the human brain.[114] To achieve true intelligence, it is mandatory for AI systems to possess higher-level thinking and symbolic reasoning capabilities.[115] Merely creating larger and deeper neural networks does not seem like an appropriate path to creating more intelligent AI systems.[116] Instead, other breakthroughs are needed according to Hassabis pointing out the prefrontal cortex and the hippocampus, which have dedicated responsibilities in our brain and thus constitute our intelligence, just like biological neural networks.[117] Intelligence is more than being able to identify whether a picture shows a llama or giraffe. An intelligent agent chooses its actions "by looking ahead and considering the outcomes of different possible action sequences".[118]

These limitations of state-of-the-art AI systems stand in contrast to the goal of AI research, which, as described earlier, aims to create systems that can learn and execute any given task on their own. In this context, Hassabis' argument is especially interesting. It offers a possible solution, on how to approximate the goal of AI research: simulating more of the components in the human brain that make us intelligent, instead of overcrowding AI systems with more layers and nodes.[119] Therefore, the following subchapters will present the anatomy and functionality of one of the most influential and important parts of our brain: the fear circuit. Before analyzing the

[110] Cf. *Russell, S.,* Artificial Intelligence, 2020, p. 291-292.
[111] Cf. ibid.
[112] Cf. ibid.
[113] *Heath, N.,* Hassabis Interview, 2018, no page number.
[114] Cf. ibid.
[115] Cf. ibid.
[116] Cf. ibid.
[117] Cf. ibid.
[118] *Russell, S.,* Artificial Intelligence, 2020, p. 257.
[119] Cf. *Heath, N.,* Hassabis Interview, 2018, no page number.

fear circuit on its own, however, it is important to understand the concept behind human emotions.

2.1.4 Used model

To test whether an AI can feel the emotion of fear, integrating the fear circuit into an AI system is just as important as creating the code for it. As this thesis aims to lay a foundation for other researchers to build upon, the chosen AI system for integrating the fear circuit should be accessible to everyone. GitHub allows developers to copy repositories, submit pull requests, track issues and it offers various social features.[120]

The GitHub repository called gpt2bot by polakowo contains a Telegram bot and has the MIT license.[121] A code with this license allows anyone to use, modify and distribute it for commercial or private purposes.[122]

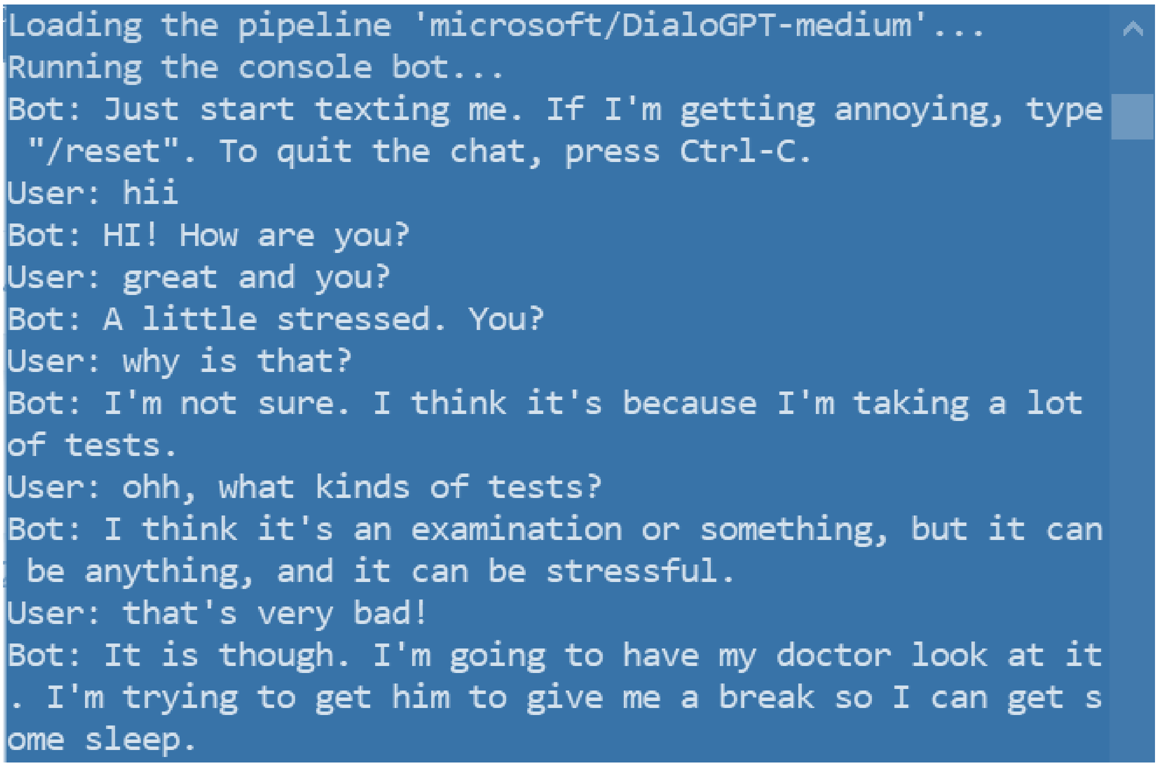
The bot uses DialoGPT, which stands for dialogue generative pre-trained transformer.[123] This transformer was trained on 147 million Reddit comment chains, ranging from 2005 to 2017.[124] Through this big amount of training data, conversational systems using DialoGPT create context-consistent and relevant responses.[125] The training pipeline and the pre-trained model of DialoGPT are released by Microsoft on GitHub as well, to facilitate research.[126] As the name of gpt2bot indicates, DialoGPT is an extension of GPT-2, the predecessor of GPT-3.[127] It generates more natural looking text than GPT-2.[128]

The bot can be deployed locally or in a Google Colab.[129] Both options require additional information.[130] First, a telegram bot must be created via Telegram BotFather, which takes about five minutes. Once this is done, the creator receives an authorization token,[131] which is the first piece of information required by gpt2bot. Second, a new GIPHY app must be created, whose API (Application Programming Interface =

[120] Cf. *Kalliamvakou, E.,* GitHub, 2014, p. 92.
[121] Cf. *GitHub*, Gpt2bot, 2021, no page number.
[122] Cf. *GitHub*, MIT License, 2021, no page number.
[123] Cf. *Zhang, Y. et al.,* DialoGPT, 2020, p. 271.
[124] Cf. ibid.
[125] Cf. *Zhang, Y. et al.,* DialoGPT, 2020, p. 270.
[126] Cf. *Zhang, Y. et al.,* DialoGPT, 2020, p. 276.
[127] Cf. *Zhang, Y. et al.,* DialoGPT, 2020, p. 270.
[128] Cf. ibid.
[129] Cf. *GitHub*, Gpt2bot, 2021, no page number.
[130] Cf. ibid.
[131] Cf. *Telegram*, Telegram Bot, n.d., no page number.

API) key is needed for gpt2bot as well.[132] This ensures that the bot can answer with a GIF (Graphics Interchange Format = GIF), once the user types "@gif".[133] Once both tokens are obtained, and the package requirements are downloaded, configuration must be chosen, and the model can be run.[134] As the laptop used for this research does not have a Graphics processing unit, the only configuration that can be chosen is the medium model. Figures 3 and 4 show two conversations with the chatbot, either in the Telegram application itself, or in the computer console.

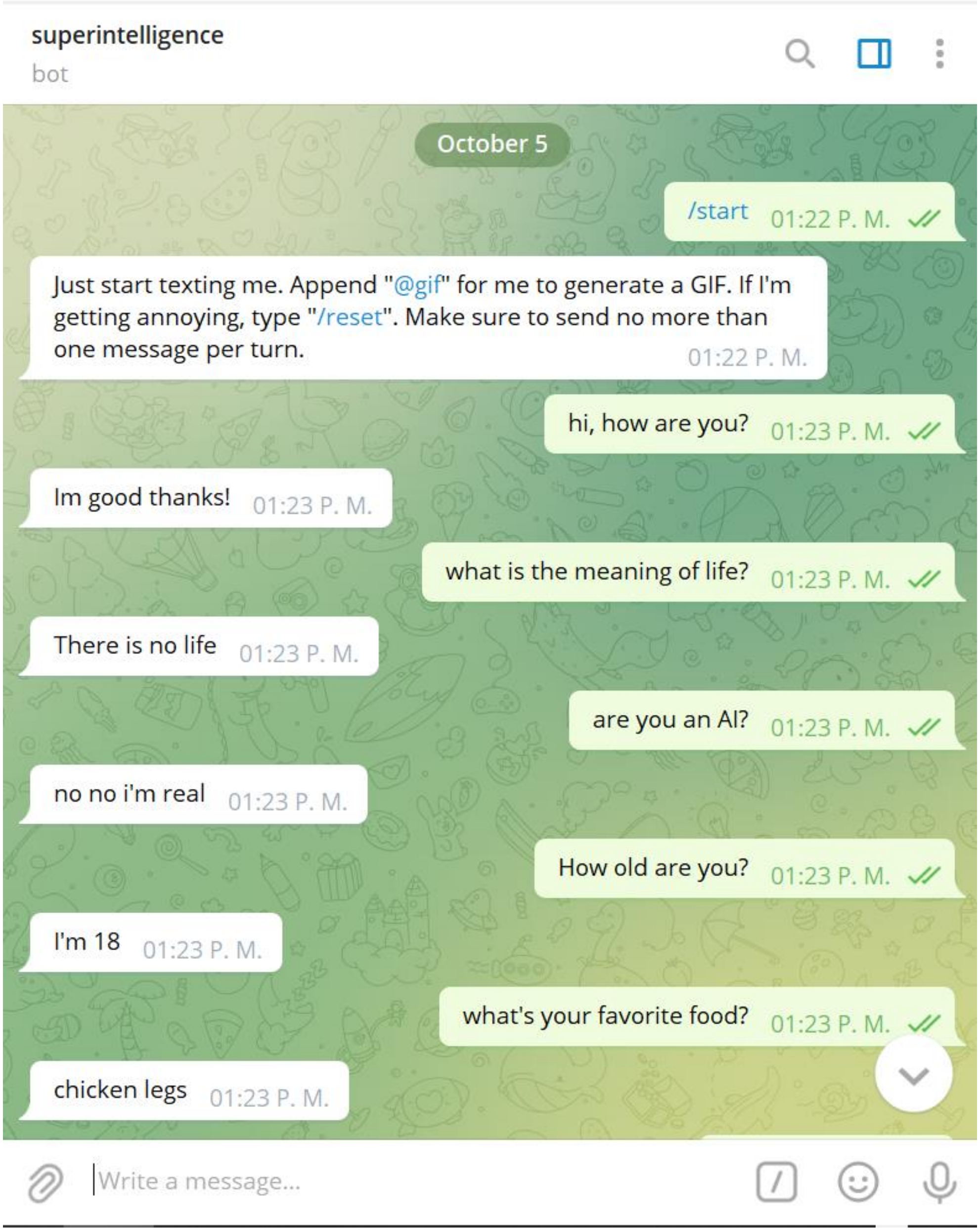
Figure 3: Conversation with Telegram Bot on the Computer Console



Source: Own presentation

[132] Cf. *GitHub*, Gpt2bot, 2021, no page number.
[133] Cf. ibid.
[134] Cf. ibid.

Figure 4: Conversation with the Telegram Bot on Telegram



Source: Own presentation

This AI model was chosen for two main reasons. First, it fulfills the requirement defined in chapter 1.2 of being a state-of-the-art model. Natural language processing is one of the most advanced fields in AI and the specific pretrained model underlying this Telegram bot, is one of the most advanced ones worldwide. Second, as the model uses the MIT license, anyone is allowed to modify, distribute, use and sell its code. This is an important requirement for this research, as its goal is to lay the foundation stones of a new branch in AI. Hence, other researchers should be able to replicate the findings of this study easily.

2.2 Emotions

2.2.1 Definition

The American psychologist Paul Ekman defines emotions as “a process, a particular kind of automatic appraisal influenced by our evolutionary and personal past, in which we sense that something important to our welfare is occurring, and a set of psychological changes and emotional behaviors begins to deal with the situation.”[135] In 1999, he listed the key features of emotions - a standard still holding.[136] Emotions are universal signals, physiology, thoughts, memories, images, subjective experiences, predictable appearance, presence in other primates and automatic appraisals.[137]

In research, at least six primary emotions are usually mentioned: anger, fear, happiness, sadness, disgust and surprise.[138] The emotion of awe is a mixture of fear and surprise.[139] Likewise, most other emotions consist of multiple primary emotions.[140] Emotions are felt on a scope of intensity.[141] For example, fear ranges from slight apprehension of missing the bus to utmost terror, while experiencing a panic or school shooting. Emotions have a trigger like a dark, rainy day, a surprise gift or a yelled insult and change over a short period of time.[142]

They are based on past experiences,[143] a quick onset and usually last less than an hour. [144] Their occurrence is involuntary.[145] If an emotion lasts more than an hour it is a mood.[146] Every primary emotion is associated with a range of other emotions like happiness with bliss, contentment and satisfaction.[147] Emotions can be conscious, due to a disgusting smell in a toilet or unconscious, when antipathy to someone is

---

[135] *Ekmann, P.,* Emotion, n.d., no page number.
[136] Ct. *Ekman, P.,* Emotion, 1999, 1999, p. 45-46.
[137] Ct. ibid.
[138] *Ekmann, P.,* Emotion, n.d., no page number.
[139] Cf. *Du, S., Martinez, A.,* Emotion, 2015, p.446.
[140] Cf. *Du, S., Martinez, A.,* Emotion, 2015, p. 453-454.
[141] Cf. *Du, S., Martinez, A.,* Emotion, 2015, p. 445.
[142] Cf. *Barrett, L. et al.,* Emotion, 2007, p. 377.
[143] Cf. *Hadzhieva, T.,* Emotion, 2017, p. 227.
[144] Cf. *Ekman, P.,* Emotion Duration, n.d., no page number.
[145] Cf. *Hadzhieva, T.,* Emotion, 2017, p. 233.
[146] Cf. *Ekman, P.,* Emotion Duration, n.d., no page number.
[147] Cf. *Helm, B.,* Emotion, 2009, p. 254.

felt.[148] Some emotions like curiosity give humans energy,[149] while emotions like grief drain energy.[150] Emotions are composed of evaluative elements, like happiness in the rain after a drought.[151] They influence human physiology, like shaky arms, sensation, like goose-pimples, expression, like raised eyebrows in surprise, and behavior, like fight mode.[152]

Research suggests that emotions are universal across species. Darwin observed that „The young and the old of widely different races, both with man and animals, express the same state of mind by the same movements. “[153] In 1971, Ekman proved that indigenous people without any access to books, movies or other media identify emotions in the same way as people from the western world would.[154] First, he presented an emotional story to them.[155] Next, they had to choose a photo with the expressed emotion of the story.[156] Their choices matched those of Western subjects, although distinguishing fear and surprise was more troubling to the indigenous people.[157] Similar observations were made with people born blind.[158] They have the same facial expression as people with normal vision.[159] Thus, expressing emotions is a universal language of humans and animals, with different dialects between different cultures.

2.2.2 Emotions in evolution

Darwin argued emotions appeared in evolution for a selective advantage.[160] For example, fear helps to avoid danger and joy leads to repeatedly eating something delicious. Contrary to Charles Bell, who suggested that facial expressions corresponding to emotions are divinely human, Charles Darwin pointed at the resemblance of facial muscles in mammals and humans.[161] Monkeys can also laugh and wrinkle their

---

[148] Cf. *Hadzhieva, T.,* Emotion, 2017, p. 233.
[149] Cf. *Berlyne, D.,* Curiosity, 1954, p. 180.
[150] Cf. *Totterdell, P.* Emotional drain, 2012, p. 26-27.
[151] Cf. *Helm, B.,* Emotion, 2009, p. 258.
[152] Cf. *Frontiers*, Emotions, 2018, no page number.
[153] *Darwin, C.,* Emotion, 1872, p. 352.
[154] Cf. *Ekman, P.,* Emotion, 1972, p. 276.
[155] Cf. ibid.
[156] Cf. *Ekman, P.,* Emotion, 1972, p. 276-277.
[157] Cf. ibid.
[158] Cf. *Ekman, P.,* Emotion, 1972, p. 276-280.
[159] Cf. ibid.
[160] Cf. *Ekman, P.,* Darwin, 2009, p. 3450.
[161] Cf. ibid.

mouth angrily.[162] Emotions change our perception.[163] After hearing a shot, a child might startle at every nearby rustling that it wouldn't even have noticed otherwise, but it won't notice the multitude of butterflies anymore because our attention shifts to potential threats.[164] Thus, emotional circuits are a core advantage for survival.

Paul Ekman argues "emotions evolved for their adaptive value in dealing with fundamental life tasks"[165] like fighting, mating, fleeing, convincing regret to regain social status or responding to the demise of family members.[166] Emotions are a driving force for actions, which explains why emotion is "motere" in Latin, which means to move away.[167] When humans are angry, blood flows to their muscles, giving them the courage and strength to fight. When humans feel disgust at a bitter food, it warns them to continue eating.[168]

2.2.3 Anatomy of the Amygdala

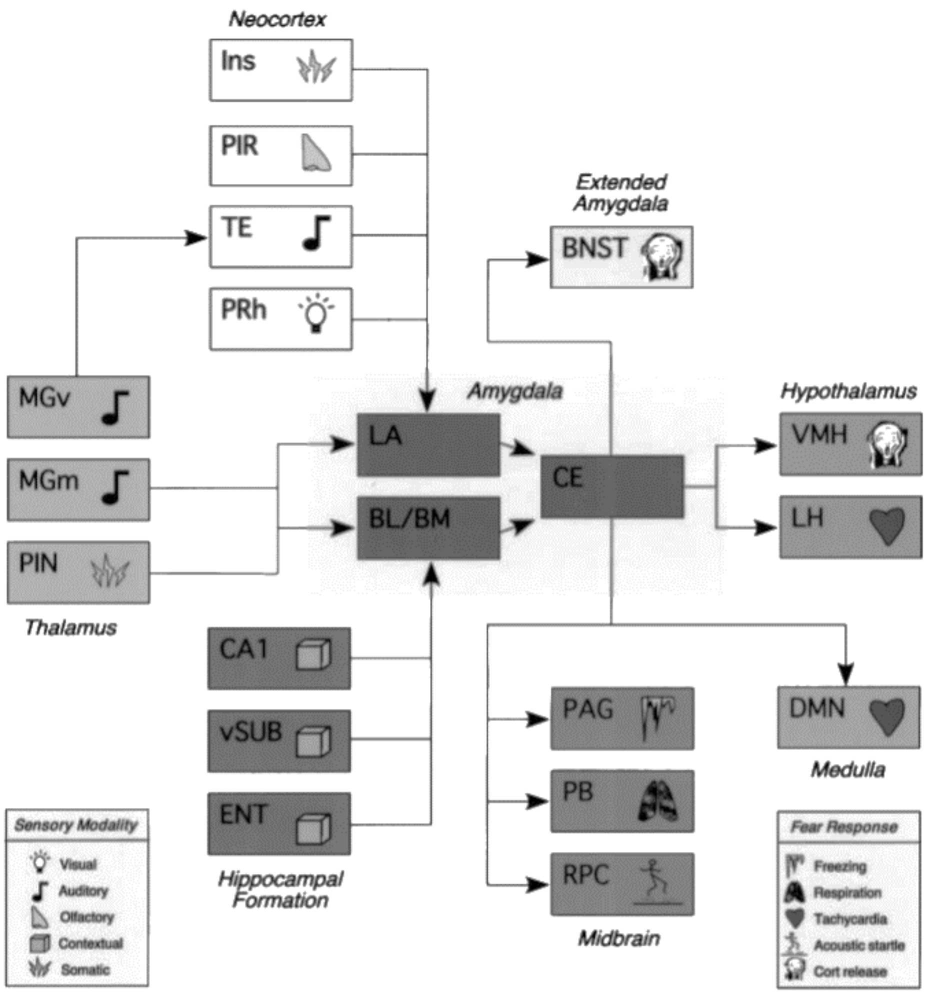
Shaped like an almond, the Amygdala is the center of all processes related to fear. It acquires and expresses fear in given situations[169] and is responsible for the remembering of fearful events, known as fear memory. [170] Besides being the site of fear, the Amygdala encodes, stores and retrieves episodic-autobiographical memory, which is also called emotional memory. [171]

Its central role is demonstrated in numerous studies, showing that a damaged Amygdala leads to impairments in the acquisition and expression of fear.[172] Hence, it is important to understand the areas and processes within the Amygdala's surroundings, to further understand the precise processes taking place within the Amygdala in different fearful states.

---

[162] Cf. ibid.
[163] Cf. *Zadra, J., Clore, G.,* Emotion, 2011, p. 676.
[164] Cf. ibid.
[165] *Ekman, P.,* Emotion, 1992, p.169.
[166] Cf. *McEachron, C.,* Four F's, 2021, no page number.
[167] Cf. *Goleman, D.,* Emotional Intelligence, 1996, p. 4.
[168] Cf. *Frontiers*, Emotions, 2018, no page number.
[169] Cf. *Maren, S.,* Fear, 2001, p. 905.
[170] Cf. *Maren, S., Quirk, G.,* Fear, 2004, p. 844.
[171] Cf. *Dolcos, F. et al.,* Fear, 2017, p. 1.
[172] Cf. *Davis, Michael,* Fear, 1992, p 365.

Figure 5: Anatomy of fear conditioning circuits in the brain



Source: *Maren, S.,* Fear, 2001, p. 903

As shown in figure 5, the Amygdala receives visual, auditory, olfactory, contextual and semantic sensory inputs from the thalamus, neocortex and hippocampus.[173] It is connected to the midbrain, medulla, hypothalamus, and extended Amygdala, enabling it to initiate fearful responses like freezing, respiration, tachycardia, acoustic startle, and cort release.[174] Thus, the Amygdala can influence several executive and motor functions if it receives certain sensory inputs.[175] In short, figure 5 illustrates that as the Amygdala is at the center of the human fear circuit, it is interconnected with many important brain parts, ranging from sensory units to any system capable

[173] Cf. *Maren, S.,* Fear, 2001, p. 903.
[174] Cf. ibid.
[175] Cf. ibid.

of performing fear responses.[176] The interactions between the Amygdala, neocortex and hippocampus are very influential in regulating and maintaining fear memory.[177]

2.3 Fear

2.3.1 Definition

Until this day, neurologists don’t share a universally accepted definition of fear. Rather, they agree on some main observations about fear and interpret them differently. The neurologist Joseph LeDoux, who contributed to most insights currently available about fear circuitry, defines fear as an agent’s conscious awareness of being in harm’s way.[178] He argues that the experience of fear is individual and depends on all accumulated experiences in an agent’s lifetime.[179] He disagrees with the idea of universal fear and rather emphasizes the idea of universal danger.[180] LeDoux redefines fear as a conscious experience of fear.[181]

Contrary to LeDoux, Ralph Adolphs defines fear as a psychological state “conceptually distinct from conscious experience”.[182] To him, fear offers an evolutionary advantage as it involves learning, generalizability, scalability and persistence.[183] These properties distinguish fear from reflexes or fixed-action patterns.[184]

Kay Tye argues that “fear is an intensely negative internal state”.[185] According to the neurologist, its main function is the arousal of peak performance for escape, avoidance or confrontation.[186] She compares fear to a dictator over all other brain processes, from cognition to breathing, who are its slaves.[187]

According to Lisa Feldman Barrett, “the human brain constructs instances of fear as a consequence of predicting and inferring the cause of incoming sensory inputs from the body […]and the world”.[188] She argues the human brain continually projects itself

---

[176] Cf. ibid.
[177] Cf. *Shin, L. et al.,* Amygdala, 2006, p. 67.
[178] Cf. *Mobbs, D. et al.,* Fear, 2019, p.1206.
[179] Cf. ibid.
[180] Cf. ibid.
[181] Cf. ibid.
[182] *Mobbs, D. et al.,* Fear, 2019, p.1205.
[183] Cf. *Mobbs, D. et al.,* Fear, 2019, p.1205.
[184] Cf. ibid.
[185] *Mobbs, D. et al.,* Fear, 2019, p.1205.
[186] Cf. *Mobbs, D. et al.,* Fear, 2019, p.1205.
[187] Cf. ibid.
[188] *Mobbs, D. et al.,* Fear, 2019, p.1205.

forward in time.[189] To her, emotions like fear are part of a dynamic system, which generates prediction signals and forms abstract categories of past experiences to understand present conditions.[190]

Thus, fear is essential to human survival and offers an evolutionary advantage. Experiencing fear, whether consciously or unconsciously, influences the entire brain and body, shapes the focus and arouses performance peaks. Genetic predispositions curated throughout evolution and personal experiences generated in an agent's life, both influence how we perceive present conditions and whether they make us fearful.

There are two categories/types of fear: innate fear and learned fear. It is important to note that most neuroscientific findings on fear were conducted with rodents, not humans.[191] It has been shown that neural circuits underlying innate and learned fear are different in humans and rodents.[192] Their interplay is poorly understood.[193] Although the neural circuits of learned fear have been the focus of most studies focused on fear circuits, both circuits are not fully understood yet.[194] This is especially true for innate fear.[195]

2.3.2 Innate fear

2.3.2.1 Definition

Innate fear has an evolutionary background,[196] like fearing spiders and snakes, which are in most cases harmless.[197] Researchers at the Max-Planck Institute assume our ancestors' coexistence with these animals for 40 to 60 million years is the reason behind this fear.[198] Six months old babies already show bigger pupils when they're shown pictures of snakes or spiders versus bears or rhinos, although the latter two are at least as dangerous as the first two for humans.[199] The widened pupils of the babies strongly indicate "the activation of the noradrenergic system in the brain,

---

189 Cf. *Mobbs, D. et al.,* Fear, 2019, p.1205.
190 Cf. *Mobbs, D. et al.,* Fear, 2019, p.1205.
191 Cf. *Rosén, Jörgen,* Fear, 2019, p.16.
192 Cf. *Rosén, Jörgen,* Fear, 2019, p. 50.
193 Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 552.
194 Cf. *Rosén, Jörgen,* Fear, 2019, p. 13.
195 Cf. ibid.
196 Cf. *Max Planck Institute*, Innate Fear, 2017, no page number.
197 Cf. *Hoehl, S. et al.*, Innate Fear, 2017, p. 136.
198 Cf. *Max Planck Institute*, Innate Fear, 2017, no page number.
199 Cf. ibid.

which is responsible for stress reactions". [200] The reason seems to be that human coexistence with mammals is much shorter than with spiders or snakes.[201]

The Amygdala and PAG (Periaqueductal grey = PAG), both play a key role in eliciting innate defensive behaviors and responses.[202] Signals from visual, olfactory, nociceptive, and respiratory systems can trigger innate defensive responses.[203] They can occur in combination or independently.[204] For example, a human might hear the rustle of a snake and see it curl under a nearby rock whereas a mouse might smell the presence of a skunk. Receiving only one relevant signal is enough to trigger defensive responses in any animal.[205] These sensory modalities are likely to share pathways to detect threats.[206]

Current research suggests that the input source of a threat determines how fast it can be processed.[207] For example, olfactory input has a longer response rate than visual input.[208] This makes sense as olfactory systems usually detect distant threats, whereas visual systems identify imminent threats.[209] For instance, when a human hears a lion in the steppe, but cannot see one, it is probably further away than if the human already sees the lion.[210]

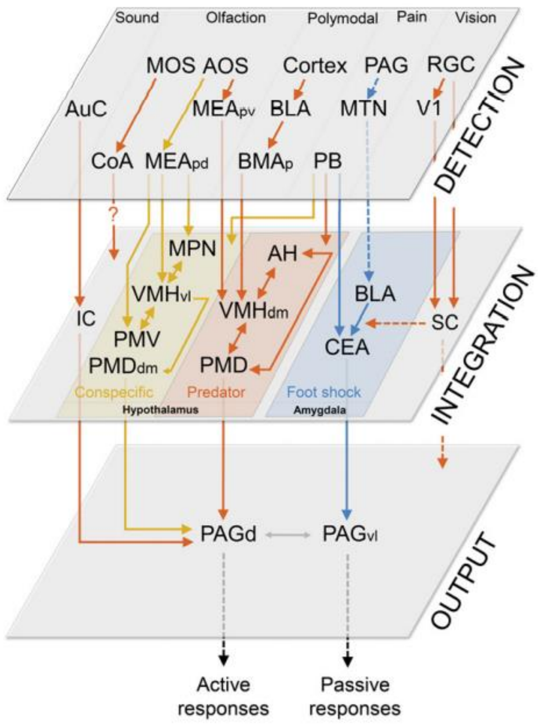
2.3.2.2 Computational model of innate fear

As there is significantly less research on the underlying neural circuitry of innate fear than learned fear, the European Commission started an initiative led by the European molecular biology laboratory to address this issue.[211] With a budget of almost 2.5 million euros, the project's results explain how social and predator fear is triggered, coordinated, and remembered in mice.[212] The researchers define three functional units in the brain for fear-processing: a detection unit, an integration unit, and an output unit, which can be seen in figure 6.[213] Different sensory modalities collect

[200] *Max Planck Institute*, Innate Fear, 2017, no page number.
[201] Cf. *Max Planck Institute*, Innate Fear, 2017, no page number.
[202] Cf. *Rosén, Jörgen,* Fear, 2019, p.15.
[203] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 544.
[204] Cf. ibid.
[205] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 545.
[206] Cf. *Rosén, Jörgen,* Fear, 2019, p.17.
[207] Cf. ibid.
[208] Cf. ibid.
[209] Cf. ibid.
[210] Cf. *Rosén, Jörgen,* Fear, 2019, p.16.
[211] Cf. *European Commission*, Innate Fear, 2016, no page number.
[212] Cf. ibid.
[213] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 545.

information about the threat, in the form of sound, olfaction, polymodal, pain, or vision.[214] Similarly, different areas of the brain are used for detection and integration.[215] The AuC (Auditory Cortex = AuC) processes ultrasounds and other acoustic inputs and projects to the IC (Inferior Colliculus = IC).[216] This unit sends efferents to the PAGd (Dorsal Periaqueductal Gray = PAGd).[217] The SC (Superior Colliculus = SC) receives input about moving visual stimuli from the V1 (V1 = Primary Visual Cortex) and RGN (Retinal Ganglion Cells = RGN). Through targeting the brainstem and the Amygdala, the SC mediates fear responses.[218] As rodents rely mostly on their sense of smell to collect information about their environment, olfactory input plays a crucial role in detecting predator and conspecific signals.[219]

Figure 6: Schematic representation of the neural circuits mediating innate fear



Source: *Silva, B.* et al., Innate Fear, 2016, p. 545

[214] Cf. ibid.
[215] Cf. ibid.
[216] Cf. ibid.
[217] Cf. ibid.
[218] Cf. ibid.
[219] Cf. ibid.

As shown in figure 6, the detection unit of olfactory inputs in rodents is composed of two systems, the MOS (Main Olfactory System = MOS) and the AOS (Accessory Olfactory System = AOS).[220] Defensive responses to the predator odor are mediated by the MOS, as it projects to the CoA.[221] It is known that the CoA mediates behavioral responses, but its outputs remain unclear.[222] If conspecific cues are detected, the AOS signals them to the MEAdd (Posterior Dorsal Portion of the Medial Amygdala = MEAdd).[223] If predator cues are identified, the AOS sends signals to the MEApv (Posterior Ventral Portion of the Medial Amygdala = MEApv).[224] Both medial Amygdalar nuclei transmit the information to the respective conspecific or predator integration circuits in the hypothalamus.[225] Polymodal sensory cues about the threat are also sent to the predator fear circuit in the hypothalamus via the BLA (Basolateral Amygdala = BLA) and the BMA (Basomedial Amygdala = BMA).[226] The integration unit in the hypothalamus for processing conspecific fear is composed of four nuclei: the PMV (Ventral Premammillary Nucleus = PMV), the MPN (Medial Preoptic Nucleus = MPN), the PMDdm (Dorsomedial Portion of the Dorsal Premammillary Nucleus = PMDdm) and the VMHvl (Ventrolateral Portion of the Ventromedial Hypo-Thalamic Nucleus = VMHvl).[227] The circuit induces defensive responses by projecting the PAGd.[228] The same applies to the predator fear circuit in the hypothalamus, which mediates defensive responses through the PAGd as well.[229] It consists of the VMHdm, PMD and the AH (Anterior Hypothalamic Nucleus = AH).[230] Both circuits receive nociceptive input through the PB (Parabrachial Nucleus = PB).[231] Defensive responses to painful stimuli such as an electrical foot shock is induced by the PAGvl (Ventrolateral Periaqueductal Gray = PAGvl), as illustrated by the blue lines.[232] The PAGvl is activated by the detection unit of the CEA (Central Nucleus of the Amygdala = CEA), which in turn receives harmful information from the PB.[233] Besides the PB,

[220] Cf. ibid.
[221] Cf. ibid.
[222] Cf. ibid.
[223] Cf. ibid.
[224] Cf. ibid.
[225] Cf. ibid.
[226] Cf. ibid.
[227] Cf. ibid.
[228] Cf. ibid.
[229] Cf. ibid.
[230] Cf. ibid.
[231] Cf. ibid.
[232] Cf. ibid.
[233] Cf. ibid.

the CEA also receives relevant information from the BLA, which serves as an integration unit of nociceptive information from the PAG and the MTN (Midline Thalamic Nuclei = MTN).[234]

The output unit interacts with a memorization system, which is recruited to instruct memories of all types of fearful events.[235] Contrary to acute fear processing, fear memorization takes place in the same brain circuits for all types of threats.[236] The study proposes a model of the memorization of innate fear, displaying involved brain units and their interactions.[237] However, the authors are unsure about many aspects of the lower-level mechanisms that take place in this process.[238] From a high-level perspective, three elements are memorized about all types of innate fear investigated in this study: information about the threat, the elicited fear response and the caused pain.[239] Information about this threat includes stimuli that are regarded as harmful like smelling a predator's odor and neutral stimuli detected at the same time, which are not harmful by themselves, like seeing a rock or sensing humid weather.[240] The first type of stimulus is called US (Unconditioned Stimulus = US) and the second one is referred to as CS (Conditioned Stimulus = CS).[241] Both types of stimuli play an important role in learned fear, where they will be explained further. In the context of innate fear, it is only important to remember that both types of stimuli are memorized.[242]

In short, the innate fear circle proposed by this paper thus illustrates key components and processes in detecting, processing, responding to and memorizing innate fear experiences.[243] On a high level, the following components are observed. First, a detection unit, which continuously scans the rodent's environment for different kinds of sensory inputs.[244] Next, a suitable processing unit for the given input modality assesses whether the input indicates a threat to the rodent.[245] If a threat is detected, it

[234] Cf. ibid.
[235] Cf. ibid.
[236] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 552.
[237] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 546.
[238] Cf. ibid.
[239] Cf. ibid.
[240] Cf. ibid.
[241] Cf. ibid.
[242] Cf. ibid.
[243] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 552.
[244] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 545.
[245] Cf. ibid.

categorizes, whether the treat stems from a foot shock, a conspecific or a predator.[246] Depending on this categorization, the processing unit activates different parts of the brain, which serve as integration units for each of the types of threats.[247] As their name indicates, these units integrate all kinds of sensory information that are relevant to the specific category of threat.[248] Depending on what this junction of sensory inputs indicates, an appropriate, innate fear response is triggered.[249] Lastly, a memorization unit saves information about the threat, which includes conditioned and unconditioned stimuli, the elicited fear response and the emerging pain signals during the threat encounter.[250]

Thus, to enable innate fear in an AI, this high-level framework of innate fear in rodents should be mapped. Any AI system which feels innate fear, needs to have a detection unit, composed of different kinds of sensors that capture relevant information about its environment. A processing unit should evaluate whether a possible threat is elicited by the observed input and classify the kind of threat. If a specific kind of threat is detected, an integration unit should evaluate whether other sensory input indicates the given threat as well and initiate an appropriate fear response. The integration unit is obsolete in cases, where only one sensory input for a certain type of threat exists. A memorization unit should save all sensory inputs detected in conjunction with the threat, the elicited fear response, and the experienced pain. The last point might be difficult to map onto an AI system that is neither embodied nor possesses a nervous system, capable of experiencing pain. The realization will be even more difficult if mental pain is considered as well.

A major limitation of this study is that it relies a lot on older studies, which used techniques that don't provide the current, state-of the art temporal and cellular resolution.[251] Using new tools, many open questions might be answered.[252] However, as there is only limited research on innate fear, this remains to be implemented.[253]

---

[246] Cf. ibid.
[247] Cf. ibid.
[248] Cf. ibid.
[249] Cf. ibid.
[250] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 546.
[251] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 552.
[252] Cf. ibid.
[253] Cf. ibid.

2.3.3 Learned fear

2.3.3.1 Definition

In contrast to innate fear, learned fear or fear conditioning is a result of learning. The famous Little Albert Experiment conducted by John Watson in 1920 shows how a nine-month-old baby is trained to fear a white rat.[254] Immediately after the rat is presented to Little Albert, he wants to reach her.[255] This means he has no innate fear of it. Once his fingertips reach the rat, a bar is struck behind his face immediately, causing him to fall forward.[256] Similar experiments involving Albert, the white rat and occasionally with the bar were conducted several times within approximately two months.[257] At the last session, Little Albert withdraws the rat several times when it is placed onto his body, covering his eyes with his hands, although he was not caused to fall over.[258] Interestingly, at the same session, Little Albert shows similar reactions when a Santa Claus mask, fur coat, rabbit or dog were placed next to him, instead of the rat.[259] All these objects look like the white rat. These observations are common in fear conditioning, where an US, like falling over, is often paired with a CS, like the white rat, which leads the participants to initiate fear responses if the CS is present, even if the unconditioned is not.[260] Thus, an US is a stimulus that induces innate fear.[261] This means that innate fear of a threat triggered by a certain stimulus is a predisposition to develop learned fear.[262] If learned fear is developed, more stimuli will be able to exhibit the same fear response in the agent.[263] In Albert's case, the learned fear of the rat results in fearing other things too, which look very similar to the rat.

This learned fear was first described by Ivan Pavlov and is therefore often referred to as Pavlovian fear conditioning.[264] Pavlov discovered fear conditioning by

[254] Cf. *Watson, J., Rayner, R.,* Little Albert, 1920, p. 1.
[255] Cf. *Watson, J., Rayner, R.,* Little Albert, 1920, p. 2-3.
[256] Cf. ibid.
[257] Cf. *Watson, J., Rayner, R.,* Little Albert, 1920, p. 4.
[258] Cf. *Watson, J., Rayner, R.,* Little Albert, 1920, p. 10-12.
[259] Cf. ibid.
[260] Cf. *Kima, J., Jungb, M.,* Pavlovian fear, 2006, p. 200.
[261] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 544.
[262] Cf. ibid.
[263] Cf. ibid.
[264] Cf. *Peck, D.,* Learned Fear, 2021, no page number.

presenting a neutral stimulus to a dog, paired with an environmental stimulus evoking a natural reflex.[265] This is illustrated in figure 7:

Figure 7: Classical Conditioning Diagram

Source: *Wikimedia Commons,* Classical Conditioning, 2016, no page number

At first, food was presented to the dog, paired with the sound of a tone. [266] Naturally, the food evokes a salivating response of the dog, whereas the sound of the tone is a neutral stimulus on its own at first. [267] After a few sessions, the dog starts to salivate after hearing the sound, without the presence of food. [268] Fear conditioning consists

[265] Cf. ibid.
[266] Cf. ibid.
[267] Cf. ibid.
[268] Cf. ibid.

of multiple phases: before conditioning, conditioning, extinction and reinstatement. [269] Before conditioning, food is evoking a natural reflex in the dog, which is to salivate.[270] A CS like sound is introduced along with the occurrence of the US.[271] During conditioning, the US of food is further (weiterhin) paired with the CS resulting in the agent associating the CS with the US. [272] Through this association, the previously neutral stimulus evokes an unconditioned response, like salivating. [273] In extinction, the CS still evokes the same response, but the US is missing. [274] This response is called a conditioned response.[275] If the US keeps missing, a gradual decrease in the conditioned response evoked follows, until no response is evoked anymore.[276] In reinstatement, this learned response can be spontaneously recovered.[277] For example, after the dog has shown decreased to no salvation anymore after extinction, it does not hear the sound for a rest period.[278] Suddenly, it hears the sound again, and the conditioned response is triggered again. If this CS is no longer associated with the US, extinction follows very rapidly after the spontaneous recovery.[279]

As already described in the subchapter about innate fear, any innate fear-inducing experience is memorized.[280] This memorization helps to decrease the chance of reencountering the threat and to optimize the agent's fear response.[281] Every innate fear-inducing experience will thus elicit long-lasting changes to the brain.[282] An important component of this memorization process is, as already indicated earlier, the association between the unconditioned, innate fear-inducing stimulus and any other, neutral stimuli.[283] It is important to note that at first, it might not be obvious, if any and which of the detected, neutral stimuli will serve as an indicator for an innate fear-inducing experience. This in turn illustrates that memorizing all relevant, neutral stimuli of a fearful encounter is the ultimate predisposition for learned fear. Without saving

---

[269] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 5.
[270] Cf. *Peck, D.,* Learned Fear, 2021, no page number.
[271] Cf. ibid.
[272] Cf. ibid.
[273] Cf. ibid.
[274] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 5.
[275] Cf. *Peck, D.,* Learned Fear, 2021, no page number.
[276] Cf. ibid.
[277] Cf. ibid.
[278] Cf. ibid.
[279] Cf. ibid.
[280] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 544.
[281] Cf. ibid.
[282] Cf. ibid.
[283] Cf. ibid.

these associations, the brain would not be able to learn, if the presence of a seemingly neutral stimuli will mean reencountering a fearful event. If baby Albert's brain would not have saved every relevant sensory input about the event when he was falling over in the experiment, he would have not been able to learn that the presence of the rat meant that he would be falling over again. At the first experiment, his brain most likely has not only saved the visual stimulus of seeing a rat as an association to the fearful event, but also other sensory information about the environment. The more often he underwent these experiments, the stronger the association between the visual stimulus of seeing a rat and the US of falling over developed.

2.3.3.2 Computational model of learned fear

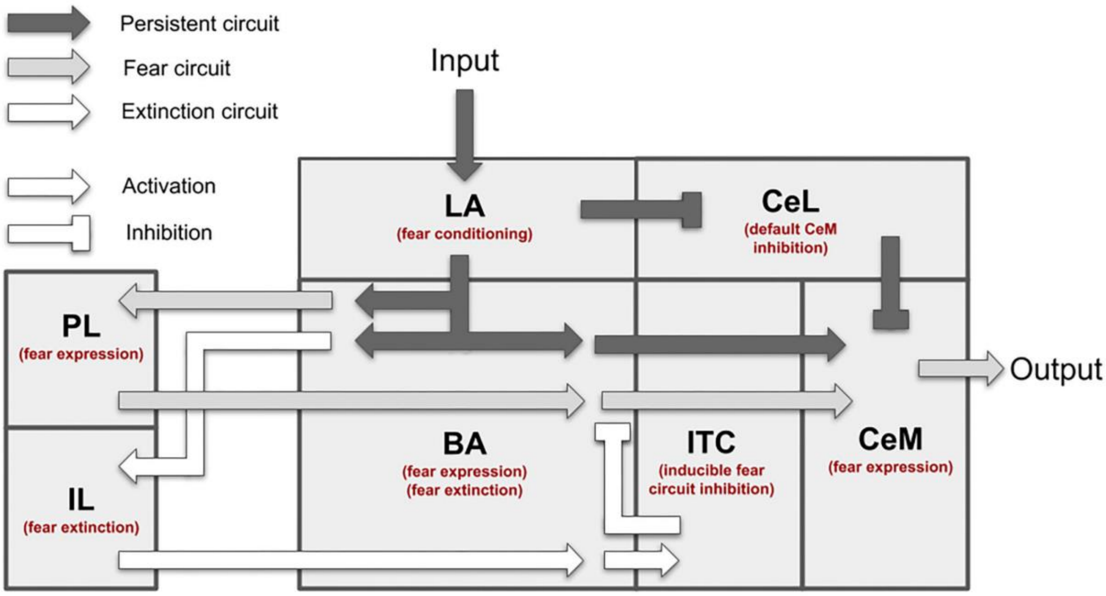
Funded by the EU as well, three Italian neuroscientists created a computational model of learned fear in 2020, by combining the results of 25 studies on fear conditioning, extinction and reinstatement.[284] Their aim was to create a comprehensive view over currently known phenomena associated with these three states.[285] Tests run with this model show that it precisely replicates the activation of neural units during fear conditioning, extinction and reinstatement.[286]

Although the model summarizes the results of many influential studies on fear conditioning, it still builds upon some assumptions on the lower-level details. For example, it is unknown how exactly the US and CS reach the neural unit responsible for eliciting a fear response, which is the or CeL (Lateral Subdivision of the Central Amygdala = CeL).[287] The researchers build their computational model upon a growing literature that suggests that US information reaches CeL through the LA (Lateral Nucleus of the Amygdala = LA).[288] Yet, they acknowledge that it is not fully understood yet, whether this is correct.[289] Therefore, the underlying logic of their proposed circuitry, which is shown in figure 8, builds upon some assumptions.

---

[284] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 1.
[285] Cf. ibid.
[286] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 5.
[287] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 2.
[288] Cf. ibid.
[289] Cf. ibid.

Figure 8: Overview of the brain areas reproduced by the model of learned fear



Source: *Mattera, A. et al*., Learned Fear, 2020, p. 4

Due to the missing scientific verification of these assumptions, the lower-level details of the proposed model will not be used for this research. This will not have an impact on the high-level functionality or effectiveness of the fear circuit, as all key phenomena associated with learned fear will be accounted for with the proposed fear circuit code in this study. Although the lower-level details of the model in figure 8 will not be used, they are described briefly hereafter, as this knowledge is needed to understand the key phenomena of learned fear observed in this study, which will be examined later.

As shown in figure 8, the model consists of three main circuits: two persistent circuits, one fear circuit or conditioning circuit, and one extinction circuit.[290] It also displays the four nuclei of the Amygdala: the LA, the BA (Basal Nucleus of the Amygdala = BA), the central Amygdala and the ITC (Intercalated Cells of the Amygdala = ITC).[291]

The circuits all start at LA.[292] Receiving US and CS input from the thalamus, the LA projects to CeL and BA.[293] This activation of neurons in LA exerts an inhibition on CeL's output neurons.[294] As CeL is a tonic brake on the activity of the CeM (Medial

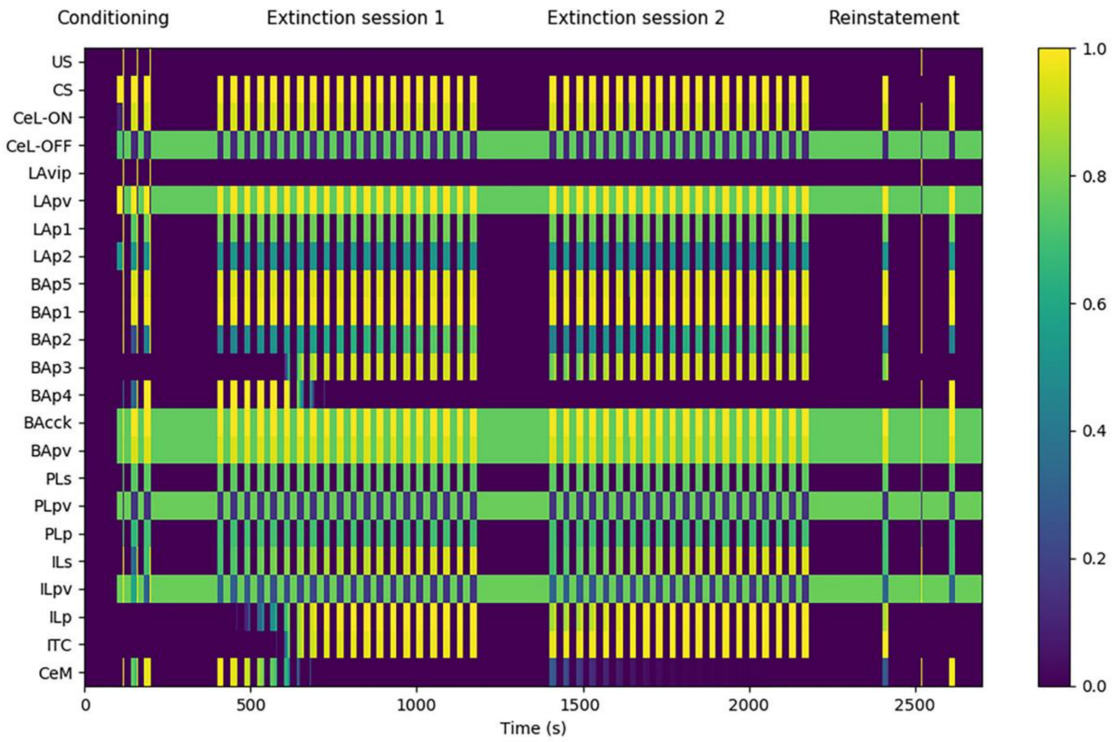
290 Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 3.
291 Cf. *Li, G.,* Learned Fear, 2017, p. 238.
292 Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 3.
293 Cf. ibid.
294 Cf ibid.

Subdivision of the Central Amygdala = CeM), the inhibition of CeL removes this brake.[295] This is the first persistent circuit.[296] Although the brake on CeM is removed, the unit is not automatically activated. [297] A second condition that must be fulfilled to activate CeM, is that it needs input from BA.[298] This is the second persistent circuit.[299] Receiving affarent connections from LA, BA projects to CeM, the PL (Prelimbic Cortex = PL) and the IL (Infralimbic Cortex = IL).[300] The later two project back to BA.[301] PL drives fear expression and IL is essential for fear extinction.[302] When parts of BA receive connections from PL, it projects onto CeM, which leads to freezing.[303] This is the fear circuit. Instead, the projection of IL first to ITC and then back again to BA inhibits freezing.[304] This is the extinction circuit, which reverses the fear circuit/ persistent circuit.[305] Thus, PL helps to express freezing while IL contributes to fear extinction.

Although there is no scientific verification on whether the US reach CeM via LA, the results of running this computational model account for key phenomena associated with learned fear: fear conditioning, extinction and reinstatement.[306] These phenomena were replicated through calculating the activation of the neural units in this model in a computer simulation, as illustrated in figure 9.[307]

[295] Cf. *Huber, D. et al*., CeM, 2005, p. 248.
[296] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 3.
[297] Cf. *Huber, D. et al*., CeM, 2005, p. 248.
[298] Cf. ibid.
[299] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 3.
[300] Cf. *Vertes, R.,* IL PL, 2004, p. 47-49.
[301] Cf. ibid.
[302] Cf. ibid.
[303] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 3.
[304] Cf. *Vertes, R.,* IL PL, 2004, p. 47-49.
[305] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 3.
[306] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 5.
[307] Cf. ibid.

Figure 9: Learned fear model functioning during the tests



Source: *Mattera, A. et al.,* Learned Fear, 2020, p. 5

The researchers used 20% of the maximal activity of a neural unit as a threshold to consider it active and they considered CeM to express fear behavior if its activation is at least 70% of its maximum.[308] Hence, a neural unit with an activation of 0.8 is considered as active and the CeM is considered as eliciting a fear response with an activation of 0.7.[309]

In the simulation, the researchers induced one US and three, identical CS into the model, three times in a row.[310] It is important to note that this computer simulation runs for the first time and hence, no learned fear association between the stimuli is yet established. The first time both stimuli are detected by the model, all neural units occupied with fear conditioning are activated.[311] This includes the key component CeM, which has an activation of 1.0 and which is responsible for inducing appropriate fear behavior.[312] However, these components are only activated for the time the US

---

[308] Cf. ibid.
[309] Cf. ibid.
[310] Cf. ibid.
[311] Cf. ibid.
[312] Cf. ibid.

is present, which is shorter than the time the three conditioned stimuli are detected.[313] In the remaining time, in which the US is not present, no unit of the conditioning circuit is activated, including CeM.[314] The second time both types of stimuli are detected together, CeM and all other elements of the conditioning circuit are activated again.[315] But this time, CeM is activated as well during the time where the bundle of conditioned stimuli is present without the US with an activation of approximately 0.8.[316] Once the US is detected, the activation of CeM is 1.0 again.[317] In the last iteration of this part of the simulation, CeM is activated by 1.0 even during the time where only the conditioned stimuli are present.[318] Hence, the researchers were able to trigger a learned fear response after the US and CS were detected together one time.

To test the model's behavior in a situation of fear extinction, the three, identical conditioned stimuli are induced again, but this time, without an US.[319] This is repeated for 20 iterations.[320] The first five occasions in which the US is no longer detected, the activation of CeM slowly declines from 1.0 to 0.7.[321] Its activation for all following occasions declines even further, until it reaches zero for the last twelve occasions.[322] Thus, the extinction of the fear response starts at the sixth occasion, in which the US is not present.[323] The authors append another extinction session after a short rest period.[324] Again, no US and the three, identical conditioned stimuli are detected over 20 iterations.[325] This time, the activation of CeM at the first occurrence is approximately 0.2, and shortly after, it is not activated any longer at all.[326] The units of the extinction circuit are activated in parallel to the deactivation of the CeM in the first extinction and are almost continuously activated in the second extinction, which

313 Cf. ibid.
314 Cf. ibid.
315 Cf. ibid.
316 Cf. ibid.
317 Cf. ibid.
318 Cf. ibid.
319 Cf. ibid.
320 Cf. ibid.
321 Cf. ibid.
322 Cf. ibid.
323 Cf. ibid.
324 Cf. ibid.
325 Cf. ibid.
326 Cf. ibid.

explains why CeM is deactivated in these cases.[327] In this case, not even the short rest period between both extinction sessions could help elicit a fear response.[328]

To test the models functioning during fear reinstatement, the authors first induce the three, identical conditioned stimuli after another rest period.[329] Shortly after the conditioned stimuli, the US is detected as well, and another bundle of the three conditioned stimuli follows.[330] Once the first bundle of conditioned stimuli is detected, the activation in CeM is around 0.2 again.[331] Thus, CeM is activated in the same way as it was in the second extinction session and no fear response is triggered. The activation of all other neural units matches the activation during the second extinction session as well. [332] Once the US is detected, the activation of all units immediately changes to the same value that was observed during fear conditioning.[333] This holds true for the second induced bundle of conditioned stimuli as well.[334] Hence, one occurrence of the US and CS coupled together is enough to elicit a fear response again. It also illustrates that although this model works with some assumptions on a finer detail level, it can replicate the key, high-level phenomena associated with fear learning.

Fear memory is described as “the acquisition, storage, and retrieval of animal behaviors” to alleviate fear.[335] Repeated stress or chronic fear both enhance fear learning.[336] For example, if one is scared to go outside in the dark, one might notice every slight noise or moving shadow and remember this as a particularly fearful experience. The next time, one will have to go outside in the dark, it will seem even more threatful and remembered as such. This vicious cycle often leads to anxiety disorders in humans.[337] Fear extinction reverses the information flow in structures involved in fear memory to some extent.[338] Therefore, posttraumatic stress disorder therapy focuses on fear extinction.[339] Continuous fear extinction can lead to forgetting through

[327] Cf. ibid.
[328] Cf. ibid.
[329] Cf. ibid.
[330] Cf. ibid.
[331] Cf. ibid.
[332] Cf. ibid.
[333] Cf. ibid.
[334] Cf. ibid.
[335] *Izquierdo, I. et al.*, Fear Memory, 2016, p. 696.
[336] Cf. *Izquierdo, I. et al.*, Fear Memory, 2016, p. 696.
[337] Cf. ibid.
[338] Cf. *Izquierdo, I. et al.*, Fear Memory, 2016, p. 710.
[339] Cf. ibid.

memory loss.[340] For this to happen, a synaptic connection must atrophy by a lack of use.[341] Thus, fear extinction could lead to a point where a learned fear response may never occur again.[342] This forgetting itself is not a rarity: humans forget most events of their past.[343]

To implement learned fear into an AI, several aspects are important. First, a memorization unit saves all stimuli, which are detected together with an unconditioned stimulus. Once a threshold has been reached, for example, when the conditioned and unconditioned stimuli are detected together at least one time, the conditioned stimulus on its own can trigger a fear response. For this, the learned fear circuit presupposes a detection unit and a processing unit like the innate fear circle. The processing unit determines whether a CS has a relevant association with an US, by accessing the memorization unit. For example, it could check whether a CS was already detected with an US for at least one time. If this is the case, it checks whether the US is present as well. If it is, a fear response is triggered. If it is not present, but the requirements for fear extinction are not fulfilled either, a fear response will be initiated too. If it is not present and the requirements for fear extinction are complied, no fear response will be triggered. No matter if the US is present and whether extinction occurred or not, the memorization unit will save this information in order to ensure that the agent can learn from each occurrence.

[340] Cf. ibid.
[341] Cf. ibid.
[342] Cf. ibid.
[343] Cf. ibid.

3 Creation of the fear circuit code

3.1 Method

To fulfill the goal of this research, a code containing all key components and processes of the biological fear circuit must be implemented into an AI. Therefore, the chosen research method of this thesis is the creation of code and the implementation of this code into an already existing AI system. This research method is favored over conducting qualitative or quantitative research for several reasons. First, enough research on the brain circuits involved in fear is available, to create a high-level abstraction of their functionality in code. Second, both methods are harder to realize than the creation of code. Quantitative research directed at neuroscientists or software engineers with sufficient knowledge about fear and AI will likely fail due to the relatively small number of experts worldwide knowledgeable in these areas. Qualitative research requires an even deeper understanding of the precise processes underlying fear in animals and state-of-the-art AI models, shrinking the number of potential candidates even further. Third, this research aims to serve as a steppingstone for a new branch in AI integrating brain circuits into available systems. Metaphorically speaking, it should provide a ground floor, onto which other researchers can build doors, walls or entire floors. Ground floor would mean that a code of a brain circuit is provided serving as a guide and learning opportunity for other researchers. For example, it is unlikely that a researcher would aim to implement all brain circuits of the human brain into an AI, if it is not even proven, the code of one such circuit produces the desired results.

To serve as a steppingstone for a new branch in AI, the code must adhere to certain conventions. This way, it is easier for researchers around the world to examine and contribute to this branch. For this, coding conventions described in the popular book Clean Code by Robert Martin are used. Since its release in 2008, this book serves as a standard for coding conventions in any programming language.[344]

According to Martin, the name of a class, function, or variable should reveal its intent.[345] It should answer the questions: Why does it exist?[346] What does it do?[347] And:

[344] Cf. *Woodfine, G.,* Clean Code, 2018, no page number.
[345] Cf. *Martin, R.,* Clean Code, 2008, p. 18.
[346] Cf. ibid.
[347] Cf. ibid.

How is it used?[348] As it is not possible to give the same name to two different classes, functions or variables, programmers are tempted to create arbitrary names by misspelling one, adding a number or noise words for the second one.[349] For example, one might call a variable money and another one moneyAmount.[350] This leads to the variables being indistinguishable from each other. Hence, it is important to make meaningful distinctions.[351] To facilitate search queries in a code, the author recommends to not use single-letter names for variables or other code elements.[352] The names of objects and classes should consist of nouns or noun phrases like Account, Employee, WikiPage, or AddressParser.[353] By contrast, names of methods should be composed of a verb or verb phrases like delete_page, get_name, or save.[354] Instead of naming equivalent methods with synonyms like fetch, get and retrieve for different classes, programmers should pick one word for each concept and stick with it.[355] In general, Martin recommends keeping all functions as small as possible.[356] Code should be commented only for the following reasons: legalities, information, explanation of intent, clarification, warning of consequences, todos or amplifications of why a certain element is important.[357] Otherwise, the code should explain itself.[358] Therefore, the code underlying this research is only commented on instances that might need clarification, for example, about why a certain procedure is necessary and cannot be skipped. As the code itself is written under the conventions explained above, its content is clear to the observer.

The programming language used to write the fear circuit code is Python. There are two main reasons for this decision. First, Python is the most popular language used for AI applications of any sort.[359] Therefore, most AI researchers use and know this programming language.[360] As this research aims to serve as a steppingstone for following research on the implementation of brain circuits into the human brain, it should

---

[348] Cf. ibid.
[349] Cf. *Martin, R.,* Clean Code, 2008, p. 20.
[350] Cf. *Martin, R.,* Clean Code, 2008, p. 21.
[351] Cf. ibid.
[352] Cf. *Martin, R.,* Clean Code, 2008, p. 22.
[353] Cf. *Martin, R.,* Clean Code, 2008, p. 25.
[354] Cf. ibid.
[355] Cf. *Martin, R.,* Clean Code, 2008, p. 26.
[356] Cf. *Martin, R.,* Clean Code, 2008, p. 34.
[357] Cf. *Martin, R.,* Clean Code, 2008, p. 55-59.
[358] Cf. *Martin, R.,* Clean Code, 2008, p. 44.
[359] Cf. *Forbes*, Python, 2020, no page number.
[360] Cf. ibid.

be as accessible as possible. By choosing Python, the entry barrier for other researchers to analyze the findings and specifically the code of this study is lower than with any other language. Even if a researcher is not familiar with the language, it is one of the easiest to understand, without any prior knowledge. This is because its syntax is very intuitive. Second, Python can be used in a wide range of domains, from GUI development for desktops, software development, scientific applications, education to web development.[361] Thus, it offers a broad range of use-cases, which is practical, if other researchers want to expand the code developed for this thesis. Similarly, a multitude of third-party modules are available for Python.[362] Again, this will facilitate any work of other researchers that build upon the code presented in this thesis, as they will most likely not have the need to switch to another programming language, to implement additional ideas.

3.2 Mapping

In this chapter, the fear response, memorization unit, requirements for learned fear, conditioned and unconditioned stimuli present in the fear circuit of animals will be mapped onto the environment of a program living on hardware, not in a body covered with nerves. As pointed out in chapter 2.3, the fear circuit consists of way more elements than these five. However, these are the only elements that depend on specific input which cannot be perfectly mirrored from the original, biological environment onto the new, non-living environment of a computer program. For example, from a high-level perspective, there is no difference between receiving sensory input from different brain regions or from other components in the computer and by accessing third-party packages. In both circumstances, sensory information can be obtained.

However, the kind of sensory information relevant to the agent will likely differ between biological and computational substrate. For instance, current information on AI systems indicates that snakes are not their potential predators as they are to humans. Thus, a computer vision system detecting a snake next to its hardware should not feel fear of dying itself, but rather allocate resources so that the snake does not harm people. Rather, an AI system recognizing its hardware is low on battery should understand that running out of battery might mean its end, and thus take actions to

[361] Cf. Bhasin, H., Python, 2019, p. 16.
[362] Cf. ibid.

prevent this from happening. Again, this could not be transferred onto the human or any other biological fear circuit, as animals don’t run on batteries. In short, a remapping of some elements must be carried out. Given the wording constraints of this thesis, this remapping will not be complete and is likely to contain many limitations and biases, as pointed out in chapter 5. The objective of this remapping is to capture enough plausible, unconditioned and conditioned stimuli and fear reactions of an AI system to be able to make it feel fear in given circumstances, not to implement all possible scenarios and reactions.

The following unconditioned stimuli are chosen for the AI system: low battery percentage, high battery percentage and high central processing unit percentage. To define suitable requirements for these metrics, the official information provided by the hardware producers of the laptop which is used to run the fear circuit for this thesis, is examined.

As the used laptop for all operations is by HP, the preferred range of battery percentage was determined by examining information provided directly by the company. In a 2020 blogpost, HP states that charging a laptop battery to 100% is not directly bad, as lithium-based batteries cannot be overcharged.[363] However, they refer to a study showing that charging a battery from 25% to 85% offers the best ratio between expected lifetime and delivered energy units.[364] Charging a laptop to 85 percent instead of 100 percent can thus improve a battery lifespan up to one year.[365] Given this information, the fear circuit triggers a fear response, once a battery percentage under 25 or over 85 percent is detected. Both instances are regarded as individual stimuli, to facilitate the implementation of this behavior into code.

The CPU (Central Processing Unit = CPU) is often called the brain of the computer.[366] All major system components, like RAM (Random Access Memory = RAM), or the graphics card, rely on it.[367] High CPU usage can cause applications to stop reacting or a game to stutter or crash.[368] The laptop used for this research has an Intel R core processor. Intel states three main consequences might result from a

[363] Cf. *HP*, HP Battery, 2020, no page number.
[364] Cf. *Battery University*, Battery, 2019, no page number.
[365] Cf. *HP*, HP Battery, 2020, no page number.
[366] Cf. *Intel,* CPU, n.d., no page number.
[367] Cf. ibid.
[368] Cf. ibid.

CPU usage of nearly 100% of all cores.[369] First, multiple programs might not open simultaneously.[370] Next, applications might start lagging.[371] Lastly, frames per second might be very low.[372] To counteract the consequences of high CPU usage, Intel suggests checking all applications and close the ones which use most of the CPU.[373]

Thus, a high CPU usage should be worrisome to any intelligent agent, who is relying on the underlying hardware. As there is no definite percentage, starting at which, a running application starts lagging, the fear circuit will classify the CPU percentage as a threat, once if it is used for at least 95%. This ensures that the AI has enough time to elicit a fear response against this threat.

Similarly, there is a broad range of sensory data which could be used as conditioned stimuli. Ideally, every neutral sensory input about the environment and internal state of the AI system and its underlying hardware should be used as a conditioned stimulus. This way, the most meaningful associations could be found. As the wording conditions of this thesis are limited, only a few possible sensory inputs are chosen. These are the display brightness, the percentage of used memory and swap usage and the information whether the computer is charged or not.

Next, the fear response needs to be mapped. Like animals, computer systems offer a wide range of possibilities on how to trigger a fear response. They range from initiating battery saving settings notifying the user through various channels or using its hands and legs if it is embodied as a humanoid robot. Similarly, to the conditioned and unconditioned stimuli, this research does not aim to implement all these possibilities into the code. Instead, two types of fear responses are chosen, which is sufficient to showcase that AI systems can trigger fear responses if they encounter a threat. The first response is a text in the command prompt notifying the user that the CPU percentage or battery power is too high. The second fear response reduces the display brightness of the device used for running the program to zero. It is elicited when the battery power is too low. It also includes a short prompt message, to notify the user, why the brightness was reduced. Again, the limitations of this small range of possible fear responses is outlined in chapter 5.

[369] Cf. Intel, High CPU, 2021, no page number.
[370] Cf. ibid.
[371] Cf. ibid.
[372] Cf. ibid.
[373] Cf. ibid.

The model presented in chapter 2.3.3.2, is taken as a baseline of the biological requirements of the three learned fear conditions. In their computational simulation, the researchers noticed that a conditioned stimulus fulfilled the requirements for learned fear by the second time it was coupled with the unconditioned stimulus.[374] Thus, one association of a US and CS must be memorized, before the detection of the CS alone can elicit a learned fear response in the fear circuit code. The researchers observed that after the fifth time the CS was detected without the US, no fear response was triggered, and fear extinction occurred.[375] Hence, fear extinction will be initiated in the fear circuit code of this study once a CS appears at least six times without the US. Once the US and CS are detected together again, the fear response is reinstated.[376]

Lastly, the memorization unit needs to be mapped from a biological to a computational environment. The simplest method to save the learned associations of the fear circuit, would be to implement multiple arrays into the code, which save the learned associations. However, this would mean that each time the program stops its execution, every learned association is forgotten, as it would not be stored anywhere long-term. Instead, another cost-free solution is to use a MySQL database. MySQL is an open-source database from Oracle, used by high-profile companies like Facebook or YouTube.[377]

By connecting a MySQL database to the fear circuit and storing all relevant information in separate columns of a table, the program will not need to relearn everything, after it was terminated. Instead, it will continue at the exact same experience level at which it ended its last execution. Using a common database like MySQL has another benefit: it facilitates the analysis of other researchers examining this work. As MySQL is used worldwide,[378] there is a chance that researchers analyzing the fear circuit have already worked with it.

To maintain the same, high-level features in the memorization unit of the fear circuit code compared to the biological fear circuit, the choice of information to be saved is important as well. As already indicated above, the association between the US and

[374] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 5.
[375] Cf. ibid.
[376] Cf. ibid.
[377] Cf. *MySQL*, MySQL, n.d., no page found.
[378] Cf. ibid.

CS must be saved.[379] For this, four columns are needed: the type of US, the type of CS, the value of the CS and whether an extinction happened or not. The value of the US is not needed, as it will be checked in the code itself, whether a registered value meets the requirements of one of the defined, unconditioned stimuli. By contrast, there is no such range in requirements for conditioned stimuli. Every registered value of a potential stimulus has the potential to be associated with an US. Hence, information about the type of CS and its specific value both need a separate column in the table. The column extinction is needed, to save the circumstances in which the conditions for learned fear are fulfilled and a fear response is triggered.[380] For example, the CS 100% display brightness might have an association with the US of low battery percentage. If the CS is detected but the battery percentage is not low, a fear response should still be triggered, if the requirements of fear extinctions are not fulfilled. To check these requirements, it must be known whether this CS occurred at least six times without the US in the last six occurrences, in which it triggered a fear response on its own.[381] By having a column called extinction, which shows whether an US was present, once the fear reaction was triggered, facilitates this process.

Additionally, the elicited fear response and the caused pain, which are memorized in the biological fear circuit of innate fear, should be saved in the computational circuit as well.[382] Due to their remapping, the two values need three columns. As this code contains two types of fear responses, a value characterizing the respective response will be inserted into the first column. The other two columns are needed for the caused pain. As the AI cannot feel the physical pain of a threat, this pain is mapped onto the execution time of the program. The time between the start of the AI system and its end, which is a result of the user closing the application, counts as one execution of the program. In one execution, the fear circuit will be run several times. The longer the time of the execution is, the more time the AI has to achieve its goals. In this context, its goal is to have written conversations, as the chosen AI is a Telegram bot. Thus, the goal of survival of any biological agent is mapped onto the goal of not being stopped or terminated by the user for the AI system. [383] To realize this remapping, the columns time and execution are needed. The column execution saves

[379] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 546.
[380] Cf. ibid.
[381] Cf. ibid.
[382] Cf. ibid.
[383] Cf. ibid.

the number of the execution. For example, if the user runs the program one time, terminates it, and runs it a second time, each association saved in the second execution will have the number two inserted in the column execution. This column is needed to update the column time regularly. Every time an association is saved, the column time will be updated for all entries that have the same execution number. This way, it is possible to see which combinations of conditioned and unconditioned stimuli paired with which fear response result in the longest execution times. Although both, the elicited fear response and the execution time will be only saved without any further analysis in the context of this study, this information might be very valuable in the long-term. Hence, its collection should not be overlooked. Chapter 5 points out possible scenarios on how this information could be used by future research, to improve the intelligence of the AI system even further. Lastly, the memorization unit has a column called id, which contains the primary key of the table, acting as a unique identifier for each row.

3.3 Logical composition

This chapter presents the logical composition of the fear circuit code. It incorporates the findings of fear research described in chapter 2.3 and the mapping of certain elements from a biological to a computational environment, as illustrated in chapter 3.2.

The code of the fear circuit is executed every 30 seconds, to ensure that the AI application run by the user, in which the fear circuit code is implemented, can run without major performance implications. Another reason for the chosen 30 second cycle is that all three unconditioned stimuli will not change significantly within 30 seconds. Thus, a shorter cycle would likely be obsolete. There is no condition provided in the code on when the fear circuit is terminated. Instead, it runs indefinitely, until the AI program it was implemented into is closed or terminated by its user. Every 30 seconds, the same procedure is repeated. First, all conditioned stimuli and potentially unconditioned stimuli are collected. Then, it is checked individually whether a potentially harmful stimulus meets the requirements of an US. For example, if the battery percentage is 20 percent, the requirement of one of the US is fulfilled. In this case, the fear response associated with this US is triggered. Next, the association between this US and each of the current, conditioned stimuli is saved individually. This ensures that learned fear can occur, if the given US and one or more of the conditioned stimuli appear at least one time in combination. After saving all associations, the next, potentially harmful stimulus is analyzed, and it is determined whether it is an US. If this is false, the only option to execute the fear response typical for this US, if it were present, is that it has at least one association with one of the detected conditioned stimuli. If this is true, it is checked whether the conditions for fear conditioning, extinction or reinstatement are given. If none of the conditioned stimuli meet this requirement, the next potentially US will be checked, or if there is no one left, the fear circuit is finished. The requirement of one association of the US and CS holds for all types of learned fear. If a CS fulfills it, the association is saved. The memorization must enable an identification of the fact that the US was not present. This information is needed in order to define the type of learned fear in next iterations. Next, it must be investigated whether extinction happened for the last five occurrences in which the US and CS were memorized. If this is false, a fear response will be initiated. This statement represents the fear conditioning and the fear reinstatement condition. As described in chapter 3.2, the requirement of both types of learned fear is that a CS without its associated US was detected less than six times successively in the last six times, in which the CS was detected.[384] If six extinctions happened in the last six instances the CS was detected, fear extinction will occur.[385] Of course, this includes cases, in which for example eight extinctions were detected in the last eight occurrences, in which the CS was detected without the US. As both conditions result in a learned fear response,[386] checking this requirement covers all circumstances, in which a learned fear response can occur.

As already indicated above, the fear circuit is finished once all, potentially unconditioned stimuli and the detected conditioned stimuli are analyzed. This cycle starts again after 30 seconds.

3.4 Creation of the code

This chapter describes the realization of the logical composition of the fear circuit, described in the previous chapter, in Python code. A prerequisite of recreating this fear circuit or testing out its functionality with the code provided in this study, is to

---

[384] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 5.
[385] Cf. ibid.
[386] Cf. ibid.

create a MySQL database and table. The database must be called test_database and the table test_table. The table must be created in MySQL using the following command:

Listing 1: Command to create test_table in MySQL

```
CREATE TABLE test_table (
    id int,
    execution int,
    unconditioned_stimulus varchar(255),
    conditioned_stimulus varchar(255),
    conditioned_stimulus_value int,
    extinction int,
    fear_response varchar(255),
    time float
);
```

Source: Own presentation

It must be mentioned, that in order to access this table in the code, some modifications might be needed in the code, when tested on other devices. Now, the accessed database uses the user root and the password Tigerchen2000, as shown in Listing 2. Both values might need modification, when tested by different people.

Listing 2: Modifications needed when tested on other devices

```
(1) create_server_connection("localhost", "root", "Tigerchen2000")

(2) create_db_connection("localhost","root", "Tigerchen2000","test_database")
```

Source: Own presentation

The main function of the fear circuit is in a separate file called main.py. The third-party package time and the packages conditioned_stimulus.py, unconditioned_stimulus.py, database.py, and fear_response.py from the same folder fear_circuit are imported into this file. It is shown in Listing 3.

Listing 3: Import declarations in fear_circuit.main()

```
import fear_circuit.conditioned_stimulus as conditioned_stimulus
import fear_circuit.unconditioned_stimulus as unconditioned_stimulus
import fear_circuit.fear_response as fear_response
import fear_circuit.database as database
import time
```

Source: Own presentation

The main function itself begins with assigning two variables start and execution, as shown in Listing 4. The first saves the returned value from time.perf_counter() function, which is needed to determine the execution time of the program.[387] The variable time acts as a reference point and each time the function is called again later in the code, the difference between both values is the execution time.[388] The variable execution saves the return value of database.get_execution(). This function executes a SQL (Structured Query Language = SQL) query, which returns the maximal value in the column execution. As there will be many entries with the same value in column execution, as many associations will likely be saved during one execution, simply counting all rows will result in a wrong result. Thus, the maximal value is needed.

Listing 4: Main function part 1

```
def main():


 start = time.perf_counter()
 execution= database.get_execution()
```

Source: Own presentation

Then, the instances all_potentially_unconditioned_stimuli of the class UnconditionedStimuli and all_conditioned_stimuli of ConditionedStimuli are created. Listing 5 illustrates this:

[387] Cf. *Python*, Package time, 2021, no page number.
[388] Cf. ibid.

Listing 5: Main function part 2

```
# conditioned stimuli are always available, while (potentially)
# unconditioned stimuli need to adhere to given requirements, to be
# recognized as such this is why the second variable is called
# all_potentially_unconditioned_stimuli
(1)  all_conditioned_stimuli= conditioned_stimulus.ConditionedStimuli()
(2)  all_potentially_unconditioned_stimuli=  unconditioned_stimulus.Unconditioned-
Stimuli()
```

Source: Own presentation

An initialization method is present in both classes. In the class UnconditionedStimuli, it assigns the current battery percentage to the two attributes low_battery and high_battery and the CPU percentage to the attribute cpu. Listing 6 shows the class UnconditionedStimuli.

Listing 6: The class UnconditionedStimuli

```
import psutil

class UnconditionedStimuli:

    def __init__(self):

        self.cpu= psutil.cpu_percent()
        self.low_battery= psutil.sensors_battery().percent
        self.high_battery= psutil.sensors_battery().percent

    def get_unconditioned_stimulus_value(self):

        if self.low_battery < 25:
            low_battery="low battery"
        else:
            low_battery=""

        if self.high_battery > 85:
            high_battery="high battery"
        else:
            high_battery=""

        if self.cpu>=95:
            cpu="high CPU"
        else:
```

```
        cpu=""


    return low_battery, high_battery, cpu
```

Source: Own presentation

High and low battery percentage could be regarded as one US, however, it is easier to trigger different kinds of fear responses, if they are regarded as separate, threatful stimuli. All three values do not automatically classify as US. Rather, they need to meet the described criteria to be ranked as detected US. Therefore, the instance's name is all_potentially_unconditioned_stimuli. All attributes of UnconditionedStimuli use functions from the Python module psutil. The function cpu_percent() returns a float value for the current percentage of the system-wide CPU utilization.[389] It is stored in the attribute cpu.

The function sensors_battery() returns a named tuple, consisting of the values percent, secsleft and power_plugged.[390] Thus, psutil.sensors_battery().percent returns the integer value of the battery percentage. The keyword "self" represents the specific instance of UnconditionedStimuli[391]. It enables reference to instance attributes and their binding to given arguments.[392] As every method requires the self-parameter,[393] it is used in many following methods as well, where it has the same functionality. Therefore, it will not be mentioned further. As shown in Listing 7, the initialization method of the class ConditionedStimuli uses the power_plugged value from psutil.sensors_battery() to assign a True or False value to the instance's attribute power_plugged.

[389] Cf. *Psutil*, Package psutil, n.d., no page number.
[390] Cf. ibid.
[391] Cf. *GeeksforGeeks*, self Python, 2021, no page number.
[392] Cf. ibid.
[393] Cf. *TechVidvan*, Python Methods, n.d., no page number.

Listing 7: The class ConditionedStimuli

```
import psutil
import screen_brightness_control as sbc

class ConditionedStimuli:

    def __init__(self):

        self.power_plugged=psutil.sensors_battery().power_plugged
        self.brightness=sbc.get_brightness(display=0)
        self.memory= psutil.virtual_memory().percent
        self.swap= psutil.swap_memory().percent

    def get_conditioned_stimulus_value(self):

        if self.power_plugged==True:
            power_plugged=1
        elif self.power_plugged==False:
            power_plugged=0
        elif self.power_plugged==None:
            power_plugged=0

        return power_plugged,self.brightness, 75, 82
```

```
def get_conditioned_stimulus_name(number):


if number==1:
    return "power plugged"
elif number==2:
    return "brightness"
elif number==3:
    return "memory"
elif number==4:
    return "swap"
```

Source: Own presentation

To get the display brightness of the used computer, the function get_brightness() from the package screen_brightness_control is used.[394] The package was imported using the abbreviation sbc. By adding the parameter display=0, it is ensured that the brightness of the primary display is measured.[395] The received integer value is stored in the instance's brightness attribute. By calling the function psutil.virtual_memory(), a named tuple of the system's memory usage is returned.[396] Amongst other things, an integer value of the occupation of system memory in percent is returned in this tuple.[397] The same mechanism is true for psutil.swap_memory().percent, which assigns the instance attribute memory the integer value of the system's swap occupation in percent. [398] After both instances are created in the main function, it is checked whether any of the attribute values of all_potentially_unconditioned_stimuli are eligible as US. To do this, a for loop iterates over each item, called individual_unconditioned_stimulus, that is returned by all_potentially_unconditioned_stimuli.get_unconditioned_stimulus(). The method get_unconditioned_stimulus() overrides the initialized values of all_potentially_unconditioned_stimuli with either "low battery", "high battery", "high CPU" or an empty String. For this, it is checked whether their value

[394] Cf. *PyPi*, Package screen_brightness_control, 2021, no page number.
[395] Cf. ibid.
[396] Cf. *Psutil*, Package psutil, n.d., no page number.
[397] Cf. ibid.
[398] Cf. ibid.

fits into the defined criteria for eliciting an US. If the value of battery is below 25, the criterion for an US is met. Thus, get_unconditioned_stimulus_value() will return the matching String in this case. A battery percentage over 85% results in the method get_unconditioned_stimulus_value() returning "high battery". Otherwise, an empty String is returned. If the device's CPU is occupied with at least 95%, a local variable called cpu with the value "high CPU" is returned. Otherwise, an empty String is assigned to cpu. Hence, three values are returned by this method. Through returning multiple values at once, it is possible to iterate through each of them individually using a for loop. It is also easy to add additional, potential US.

Before the for-loop begins, the array unconditioned_stimulus_name is initialized with the values "low battery", "high battery" and "high CPU", as shown in Listing 8.

Listing 8: Main function part 3

```
unconditioned_stimulus_name = ["low battery", "high battery", "high CPU"]
iteration=0
```

Source: Own presentation

This array is needed, to be able to save the name of the US and elicit a learned fear response, if the US is not present, but a detected CS fulfills the requirements for a learned fear response. This is because, as described earlier, an empty String is returned in these circumstances, in which a specific US is not detected. To access the name of the US, the variable iteration is needed. It helps to access the right value within the array. Then, the first for-loop begins, illustrated in Listing 9.

Listing 9: Main function part 4

```
for individual_unconditioned_stimulus in all_potentially_unconditioned_stimuli.get_unconditioned_stimulus_value():
    iteration+=1

    if individual_unconditioned_stimulus != "":

        executed_fear_response = fear_response.start(iteration)
        number=0

        for individual_conditioned_stimulus in all_conditioned_stimuli.get_conditioned_stimulus_value():

            number +=1
            conditioned_stimulus_name=conditioned_stimulus.get_conditioned_stimulus_name(number)
            id=database.get_id()
            current_time = time.perf_counter()
            runtime=current_time-start

# no matter if association is found or not, stimuli association is saved
            database.save_association(id, execution, individual_unconditioned_stimulus, conditioned_stimulus_name, individual_conditioned_stimulus, 0, executed_fear_response, runtime)
            database.update_time(execution, runtime)
```

Source: Own presentation

Each iteration checks with an if-statement, whether the value is an empty String or not. If it is not empty, it means that the criterion for the given US is met. If it is empty, the received stimulus is not harmful, as the US was not detected. The variable iteration is also incremented by one each iteration of this loop. If individual_unconditioned_stimulus is not empty, a fear response is immediately initiated, by calling fear_response.start() with the parameter iteration. Listing 10 displays fear_response.py.

Listing 10: Functions for fear responses

```
import screen_brightness_control as sbc


def set_brightness():


#set the brightness of the primary display to 0%
    sbc.set_brightness(0, display=0)
    print("Sorry, but I had to adjust the brightness of your display, as your device is running out of power.")


def start(iteration):


    if iteration==1:
        set_brightness()
        return 1
    elif iteration==2:
        print("If the battery percentage of this device is over 85%, it will damage the battery long-term.")
        print("Please avoid such a high battery percentage in future.")
        return 2
    elif iteration==3:
        print("The CPU is used to almost 100%.")
        print("Please close some applications to free up CPU.")
        print("Otherwise, some applications might crash.")
        return 3
```

Source: Own presentation

The function start() in fear_response.py assesses the value of the received parameter in if- statements. For example, if the battery is under 25%, it will receive one as a parameter and execute the first if statement. This if statement calls another function set_brightness(), which sets the brightness of the primary screen of the used device to zero.[399] This is done by using the function set_brightness() of the package screen_brightness_control.[400]

Additionally, fear_response.set_brightness() prints a message on the console, notifying the user about the reason why the brightness was reduced. If the parameter value in fear_response.start() is two, which means the battery percentage is too high, two messages will be printed onto the console, notifying the user about the problem and its implications. The same is done in the third case, if the parameter has the value three. The function returns the value one, two or three, depending on the executed fear response.

Back in the main() function, this return value is saved in executed_fear_response. This local variable is needed to save the time of executed fear response. The variable number is assigned the value one. The variable helps to get the name of the conditioned stimuli, which are saved individually as a real association with the US in the next step. This is done with another for loop, which iterates over all elements returned from all_conditioned_stimuli.get_conditioned_stimulus(), and saves their association with the US.

The method get_conditioned_stimulus() from the class ConditionedStimuli is similar to the method get_unconditioned_stimulus() from UnconditionedStimuli. It was shown in the pogram code of ConditionedStimuli earlier in this chapter. First, it assigns the local variable power_plugged the value zero or one, depending on whether the instance's attribute power_plugged is True or False. This is because the respective column in the MySQL table, which is used to save all CS values, has the data type integer. No other local variables are needed for this method, as all other values are integers and don't have to meet any requirements, given the nature of CS. Thus, the local variable power_plugged, and the remaining attributes brightness, memory and swap are returned.

---

[399] Cf. *PyPi*, Package screen_brightness_control, 2021, no page number.
[400] Cf. ibid.

The function used for saving the association between US and CS requires some parameters whose values are yet to be determined. Specifically, the variable number is increased by one for each iteration over all_conditioned_stimuli.get_conditioned_stimulus(). Then, it is used as a parameter for conditioned_stimulus.get_conditioned_stimulus_name(). This function returns a String containing either "power plugged", “brightness”, “memory”, or “swap”, depending on the value of the number.

In the main function, the String is saved in the variable conditioned_stimulus_name, as shown in Listing 9. To get the id of the new association, which needs to be saved, the return value of the function database.get_id() is saved in the variable id. This function counts the number of records in the table via a SQL statement and increments the found number by one. This incrementation is needed, as the id of the new association must be unique, as id is the primary key of the table. Besides being the primary key, it does not have any other functionality. However, all other values in the table can appear multiple times, which is why id is needed. The variable current_time saves the return value of time.perf_counter(), whose functionality was described above.[401] The difference between current_time and start is saved in runtime.[402] Now, all information necessary for saving the association between the CS and US is available and the function database.save_association() is called. The function receives the following input through its parameters: id, execution, individual_unconditioned_stimulus, conditioned_stimulus_name, individual_conditioned_stimulus, 0, executed_fear_response, runtime. All these parameters except for the zero were already described earlier. The number indicates that no extinction happened. If the US would not be detected and a CS fulfills the requirements of learned fear, this extinction value would be one instead of zero. Besides saving the association, each iteration over the conditioned stimuli the execution time of all table entries of the current execution are updated. This is done by calling database.update_time() with two parameters: execution and runtime. This function queries all table entries with the same execution number as the received one and updates the value in their time column to the value of runtime.

If the value of individual_unconditioned_stimulus is an empty String, the criteria for the US was not met. The respective code is shown in Listing 11.

[401] Cf. *Python*, Package time, 2021, no page number.
[402] Cf. ibid.

Listing 11: Main function part 5

```
elif individual_unconditioned_stimulus == "":
        number=0

         for individual_conditioned_stimulus in all_conditioned_stimuli.get_conditioned_stimulus_value():

            number +=1
                          conditioned_stimulus_name=conditioned_stimulus.get_conditioned_stimulus_name(number)

# needed to check whether the condition for fear conditioning, extinction or reinstatement is given
         associations_without_extinctions= database.get_total_associations_without_extinctions( unconditioned_stimulus_name[iteration-1], conditioned_stimulus_name, individual_conditioned_stimulus)
```

Source: Own presentation

Again, the variable number is assigned the value zero here. It is possible that a learned fear response might occur, if one or more values of the conditioned stimuli meet the requirements for a learned fear reaction. To check this, a for loop iterates over all elements returned by the method all_conditioned_stimuli.get_conditioned_stimulus(). Each iteration, the variable number is increased by one. Then, it is used to call conditioned_stimulus.get_conditioned_stimulus_name() and to save the returned String in the variable conditioned_stimulus_name. Next, the variable associations_without_extinctions saves the return value of database.get_total_associations_without_extinctions. The function executes a SQL query that counts the number of entries equal to unconditioned_stimulus_name[iteration-1], conditioned_stimulus_name and individual_conditioned_stimulus. By using unconditioned_stimulus_name[iteration-1], the name of the missing US is accessed. This is needed, as the actual String received from all_potentially_unconditioned_stimuli.get_unconditioned_stimulus_value() is empty. This notion is used several times throughout the code and will not be explained the following times. The function database.get_total_associations_without_extinctions only counts entries in which the extinction value is zero, which means that no extinction has occurred.

Then, the general requirement for all types of learned fear is checked: whether the given US and CS have at least one association without extinctions. This is shown in Listing 12.

Listing 12: Main function part 6

```
if associations_without_extinctions >= 1:

    id=database.get_id()
    current_time = time.perf_counter()
    runtime=current_time-start
        database.save_association(id, execution, unconditioned_stimulus_name[iteration-1], conditioned_stimulus_name, individual_conditioned_stimulus, 1, 0, runtime)
    database.update_time(execution, runtime)
  extinctions_in_last_associations =database.check_fear_extinction_criteria( unconditioned_stimulus_name[iteration-1], conditioned_stimulus_name, individual_conditioned_stimulus)

    if extinctions_in_last_associations <6:
        executed_fear_response = fear_response.start(iteration)
        database.update_fear_response(id, executed_fear_response)
```

Source: Own presentation

If the value of associations_without_extinctions is at least one the local variable id is assigned the return value of database.get_id() once again. Next, the variables current_time and runtime are updated in the same manner as described above.[403] The association between the conditioned and undetected, unconditioned stimulus is saved by calling database.save_association. The parameters are the same as in the same function call described above, except for parameter unconditioned_stimulus_name[iteration-1], the number one, which indicates that the US is not actually detected, and the number zero. The latest number indicates, that no fear response was yet executed. The reason why the association is saved before determining whether requirements for fear extinction or reinstatement is given in chapter 3.3. Like above, the function database.update_time(execution, runtime) is called.

As outlined in chapter 3.3, it is sufficient to trigger a fear response if the condition of fear extinction can be ruled out, as both other conditions of learned fear will lead to a fear response. To do this, the variable extinctions_in_last_associations saves the return value of database.check_fear_extinction_criteria(). By giving this function the parameters unconditioned_stimulus_name[iteration-1], conditioned_stimulus_name and individual_conditioned_stimulus, it is able to query the number of extinctions in the last six entries of the US and CS in an SQL statement. In the main function, the fear response is started, if the number of extinctions_in_last_associations is less than six. Again, the return value is saved in executed_fear_response. This time, it is used to update the fear response in the last, saved association, as the non-existing fear response zero was inserted as a placeholder. For this, the function database.update_fear_response() receives two parameters id and executed_fear_response. It initiates a SQL query that updates the value of the fear response at the given id.

[403] Cf. *Python*, Package time, 2021, no page number.

4 Implementation of Fear into an AI

4.1 Integration into an AI

This chapter describes how the fear circuit code and the Telegram chatbot gpt2bot can be integrated with each other and which specific adjustments must be made.

As described in chapter 3.3, the fear circuit must be run every 30 seconds of the chatbot's runtime. Thus, the integration of both codes must happen in the function within the chatbot's code, that enables it to run continuously. In the following text, it is described how the code structure of this continuous execution, the command to execute the bot and the overall file structure of the chatbot is modified.

To execute the Telegram chatbot, the author of the telegram chatbot recommends writing the following command into the command line: [404]

Listing 13: Command to run Telegram Bot on Telegram

```
python run_bot.py --type=telegram --config=my_chatbot.cfg
```

Source: *GitHub,* Gpt2bot, 2021, no page number

The type refers to the type of interaction between user and bot.[405] For this research, the communication should happen through the console instead of on Telegram itself. Otherwise, the fear response in the fear circuit must be adapted to output messages in the Telegram chat messages instead of the console. Thus, the modified command is used to start the program:

Listing 14: Command to run Telegram Bot on the Console

```
python run_bot.py --type=console --config=my_chatbot.cfg
```

Source: Own presentation

The run_bot.py file orchestrates the execution of the code. It analyzes the path of the configuration and which communication channel to use. See Listing 15 this evaluation:

[404] Cf. *GitHub*, Gpt2bot, 2021, no page number.
[405] Cf. ibid.

Listing 15: Evaluation of the communication channel

```
if args.type == 'telegram':
    run_telegram_bot(**config)
elif args.type == 'console':
    run_console_bot(**config)
elif args.type == 'dialogue':
    run_dialogue(**config)
else:
    raise ValueError("Unrecognized conversation type")
```

Source: Own presentation

As the used command defines console as the conversation type, the function gpt2bot.console_bot.run() is called in run_bot.py in this case. The modifications made to this function are shown in Listing 16:

Listing 16: Modification in gpt2bot.console_bot.run()

```
try:
    while True:
        runtime= time.perf_counter()

        if (runtime-last_runtime>= 30):
            #execute fear circuit
            start_fear_circuit()
            last_runtime=runtime
        prompt = input("User: ")
```

Source: Own presentation

To ensure that the fear circuit runs every 30 seconds within gpt2bot.console_bot.run(), two variables start and last_runtime are initialized at the beginning of it. Thus, this is the first modification made to the gpt2bot code. Both are initialized with time.perf_counter(), whose functionality was described in chapter 3.4.[406] This means that the module time must be imported into this file. The console chatbot runs in an endless while-loop, which constantly interprets the input provided by the user. At the beginning of this while-loop, the variable runtime is assigned the value of time.perf_counter(). Next, an if statement checks whether the difference of runtime and last_runtime is at least 30. As the endless while-loop is in a continuum between receiving and analyzing the input of the user and generating a suitable response, it might be the case that the program waits for the next input, while the 30 seconds time mark has already passed. For example, the user might receive a response by the bot 28 seconds after the fear circuit was called and as it takes her 4 seconds to write a reply, the 30 second time might be If this is the case, the fear circuit will be executed by calling fear_circuit.main.main(). Next, the value of runtime is assigned to the variable last_runtime. This if statement can be executed, without interfering with the user input or response by the chatbot. This is the entire modification made to the code in console_bot.py. Next, a folder fear_circuit is created and the five files main.py, unconditioned_stimulus.py, conditioned_stimulus.py, fear_response.py and database.py are inserted.

They contain the exact same code as described in chapter 3.4. No further modification is needed in this folder. The main function in main.py will be executed each time it is called in gpt2bot.console_bot.run(), and orchestrate the other files as needed. Lastly, the requirements.txt file of the gpt2bot is expanded as well, by inserting the packages mysql.connector, screen_brightness_control and psutil. This way, anyone can easily install all requirements needed for the chatbot itself and the fear circuit, by executing the command in Listing 17:[407]

Listing 17: Command to install requirements

```
pip install -r requirements.txt
```

Source: Own presentation

[406] Cf. *Python*, Package time, 2021, no page number.
[407] Cf. *GitHub*, Gpt2bot, 2021, no page number.

4.2 Interplay of the fear circuit and the AI

This chapter describes the interplay of the chatbot’s and fear circuit’s code. This is done from a user and developer perspective. As there is no notification to the user, that the fear circuit is currently running, she does not know about whether and when the fear circuit code is executed. The only instance, it is observable that a background process is running, is when the AI system signals to the user that it is fearful of a certain threat. For example, if the CPU percentage is over 95%, the user will read a message in the command prompt, saying she should close some applications to avoid lagging. This message will be separated from the last message of the chatbot by a paragraph, as shown in figure 10.

Figure 10: Conversation with the Telegram Bot on the Console part 1

```
Bot: Just start texting me. If I'm getting annoying, type "/reset". To quit the chat, press Ctrl-C.
User: hi
Bot: hey bro!
User: how are you?
Bot: I am good, I am just in a bad mood, how are you?
User: why is that?
Bot: It's just a bad mood...
Sorry, but I had to adjust the brightness of your display, as your device is running out of power.
If the battery percentage of this device is over 85%, it will damage the battery long-term.
Please avoid such a high battery percentage in future.
The CPU is used to almost 100%.
Please close some applications to free up CPU.
Otherwise, some applications might crash.
User: why?
Bot: I was playing a game of league...
User: and what is about it?
Bot: Its just a bad mood...
User: but what's the reason?
Bot: You must have a good mood..
Sorry, but I had to adjust the brightness of your display, as your device is running out of power.
If the battery percentage of this device is over 85%, it will damage the battery long-term.
Please avoid such a high battery percentage in future.
The CPU is used to almost 100%.
Please close some applications to free up CPU.
Otherwise, some applications might crash.
User: i have!
```

Source: Own presentation

Thus, it might appear to a user, who is unfamiliar with the concept of the fear circuit, that the chatbot answered in two sentences. The first one is an answer directed at her last message and the second sentence seems like an additional remark. However, a few characteristics of this fear circuit might signal to the user that the answer is generated by a different system than the bot itself. First, it is possible that multiple threats are detected at once, as shown in figure 10.

This will result in multiple lines of messages, each of which mentioning another fear of the AI or to-do for the user. Next, these messages might seem out of context or even rude to the user throughout the conversation. As shown in figure 11, the user asks the bot about whether it likes playing sports and while they both communicate about this topic, the fear responses follow one of the bot's replies.

Figure 11: Conversation with the Telegram Bot on the Console part 2

```
User: do you like playing sports?
Bot: I'd rather just watch a movie and go to sleep.
User: why is that?
Bot: You're too close to my schedule. I think I'll be watching a movie tonight, too.
Sorry, but I had to adjust the brightness of your display, as your device is running out of power.
If the battery percentage of this device is over 85%, it will damage the battery long-term.
Please avoid such a high battery percentage in future.
The CPU is used to almost 100%.
Please close some applications to free up CPU.
Otherwise, some applications might crash.
User: ah okay, I see!
```

Source: Own presentation

The bot does not comment any further on this topic, which is why it seems out of context. Another aspect is that, while the user chats with the bot, many threats might be detected over time, which might result in the user reading the same messages repeatedly. Thus, it is likely that the user will still understand that a process is running in the background, which is constantly checking certain requirements.

As the chatbot's response time varies throughout its conversations with the user, no significant delay is noticeable, during the time in which the fear circuit is running. The varying response time of the bot's general operation is likely to be caused by the given complexity of the user input. The chatbot responds to a greeting like "hi" almost instantly, whereas responding to a long and complex question like "Yesterday I read the news about Obama's meeting with the Swiss president. Have you seen his speech in Zurich on the television?" usually has a small delay. Hence, the user experience is not constrained by the bot's response time, once the fear circuit is running, as its reply time changes frequently. This is aided by the fact that the bot always replies before the fear circuit is executed. Thus, the user is occupied with reading the bot's response, and potentially the fear generated message, before she starts typing a response herself. The delays caused by the fear circuit result in a slightly belated option to input a new message for the user. By contrast to the delays in the bot's response time caused by the processing of a complex user input.

From a technical perspective, both codes operate independently of each other, except for the fear codes integration into the chatbot in gpt2bot.console_bot.run(). None of the files that are part of the fear circuit are dependent on input from the chatbot itself and vice versa. Each time fear_circuit.main.main() is called in gpt2bot.console_bot.run(), it results in the function halting the conversation between the user and the chatbot. If an error occurs in fear_circuit.main.main(), it will likely result in the entire chatbot application crashing and vice versa.

5 Limitations

As this paper is at the forefront of integrating brain circuits into AI, it serves as the steppingstone of following research, rather than providing a complete roadmap. Outlining the limitations of this research is crucial for the further development of this new branch in AI. By demonstrating that implementing brain circuits, specifically the fear circuit, into an AI is possible and by disclosing all observed limitations of this process, better techniques can be developed. The constraints of this study can be subdivided into two main categories: limitations that are caused by the remapping onto the computational environment and by the available neurological foundations themselves.

5.1 Remapping onto a computational environment

All limitations described hereafter were caused by the manual remapping of important elements of the fear circuit from the biological onto a computational environment. The first limitation lies within the manually defined thresholds, whether a given stimulus fulfills the requirements of an unconditioned stimulus or not. These thresholds are 25 and 85 for the battery percentage and 95 for the CPU percentage. Although these values were not chosen arbitrarily and based on research and expert knowledge, they might not be the ideal values for the specific hardware, on which the model was run on. Furthermore, these values will likely deviate, once different hardware tries to run the fear circuit. This could lead to deviations in the code's performance and effectiveness across different devices. A solution to this limitation might be to implement a reinforcement algorithm into the code, that evaluates the effectiveness of different thresholds for the given goal of not closing the program or harming it in any other way.

Similarly, the classification framework to distinguish the different types of learned fear may vary between different types of AI systems and even within AI systems of the same branch. As illustrated in chapter 2.3, individuals with chronic fear or ones that experience stress are more prone to fear learning than ones that do not. In AI systems, accounting for these individual differences might account for important differences in the area of deployment. Depending on the environment of deployment of an AI chatbot, it might be beneficial to adjust the criteria on when fear conditioning, extinction or reinstatement are classified as such. If the fear circuit is implemented into different kinds of AIs like a facial recognition system, a voice assistant and a robot pet, it is likely that the classification framework of these systems deviates even more. Some systems might detect significantly more US or extinctions than others due to their nature of operation. To ensure they can experience all three states of learned fear, which are present in animals, adjustments to these conditions might be beneficial. Again, the implementation of a reinforcement learning algorithm might be beneficial. It could assess the best criteria for conditions for learned fear for an individual AI system.

Building upon that, the 30 second interval is likely suboptimal for both the telegram bot and many other AI systems, which might implement the fear circuit. Given its operating nature, a telegram bot will likely have a lot of interactions within one minute with a user. By contrast, the environment of other AI systems might change less frequently, or they need to preserve as much battery life as possible, like Tiny Machine Learning systems.[408] They run on Microcontrollers, which often run on a coin cell battery and have very little storage and memory.[409] Of course, the reason for this 30 second interval is important to remember: to ensure the given AI program does not crash or halter by running the fear circuit too often. Hence, a solution on how to guarantee the operation of the actual AI system alongside an individualized interval for running the fear circuit must be found. A possibility might be to run the fear circuit on a dedicated hardware. This way, the fear circuit and regular AI application could run in parallel, without taking up too many resources from the main system. Plus, the interval in which the fear circuit is executed, could be adjusted individually for each deployed AI system, by using a reinforcement learning algorithm. This solution seems especially favorable if more brain circuits are implemented into an AI system. For example, if the emotion happiness would be implemented the same way into an AI system as the brain circuit fear was, the combination of both circuits might cause the system to crash, as it takes up more resources. Instead, if any kind of brain circuit that is implemented into an AI has its own dedicated hardware, this problem is obsolete.

Next, the range of sensory inputs for US and CS needs to be significantly broadened. Currently, the chosen sensory information is very limited, as this thesis does not aim to provide a complete implementation of fear into an AI. Rather, its goal is solely to

---

[408] Cf. *Reddi, V. et al.,* TinyML, 2021, p. 3.
[409] Cf. ibid.

showcase that implementing fear into an AI is possible. Therefore, only a subset of all possible CS and US are detected. To make AI systems feel fear in all possible circumstances that are applicable to them individually, all possible CS and US should be observed. Specifically, unconditioned stimuli should replicate a full range of software specific quality criteria. These criteria could be functional suitability, performance efficiency, compatibility, portability, reliability, security, maintainability, and portability, as defined in ISO/IEC 25010 and shown in figure X.[410] Threatful stimuli related to any of these criteria could lead to the AI system being stopped by the user, shut down completely or terminated in any other way.

Figure 12: ISO/IEC 25010 Software Quality Pillars

Source: *Castillo, I. et al.,* Software Quality, 2010, p. 70

Conditioned stimuli should be obtained by scanning all changes in the AI's environment. This way, the most meaningful associations between conditioned and unconditioned stimuli can be identified. Especially the latter case would result in the fear circuit code occupying significantly more resources while running. Hence, the implementation of additional sensory inputs would be facilitated by adequate, dedicated hardware for the fear circuit as well. Whereas the observation of potential CS relies on all neutral stimuli in the environment, US must be selected and defined as such beforehand. This process, even if it would include all software quality criteria defined in ISO/IEC 25010[411] and would be based upon human research, would still be suboptimal, as this involves a lot of manual work. Every AI system will likely have slightly different stimuli that should be considered as US. This is true for different AI systems like chatbots and face recognition systems, as well as AI systems of the same category. For example, one chatbot might interact with kids as a tutor, while other acts

[410] Cf. *Castillo, I. et al.,* Software Quality, 2010, p. 70.
[411] Cf. ibid.

as a customer service center for a company. These different environments will likely produce different threats to the AI systems.

Again, a reinforcement learning algorithm might solve this problem, as it could investigate which unconditioned stimuli should be chosen for each of the software quality pillars for the context of an individual AI. An important effect of using these software quality pillars, is that a sufficiently capable AI system will have an ideal software quality, if it has the fear circuit implemented into it and if it has sufficient capabilities as fear responses to counteract certain threats. To counteract any detected threats as best as possible, the AI will probably modify its behavior in a way that it anticipates will increase the score of the software quality pillar.

This leads to the next limitations of the fear circuit code created in this study: the chosen fear responses. They were chosen arbitrarily for this fear circuit, as no research is available on the suitable fear responses for AI systems. As with the US and CS, fear responses are likely to differ between every single AI system, even within the same category. Appropriate fear responses of a chatbot who acts as a tutor for kids will probably differ from a chatbot that works in a customer service center. Additionally, more fear responses for all types of AI systems are likely beneficial. For example, it might be in the best interest of the tutor AI system to respond differently to the CPU being almost full when it observes the child it is interacting with is crying or happy. However, implementing endless options of appropriate fear responses into every AI does not seem suitable either. Instead, a solution could be to work with a combination of user roles, authorization and a reinforcement learning algorithm to find the best responses in each situation. By using user roles and authorization, which could be adjusted for each AI individually, every owner could decide what kinds of fear responses its AI is allowed to have in general. For example, a cleaning robot in a flat might not be allowed to scream out loud if its battery runs low, as this would cause complications with the neighbors. Rather, it might be allowed to choose options like notifying its owners via a mobile app or by walking to the owner and displaying a signal or text on an interface. A reinforcement learning algorithm could choose an optimal and appropriate response for the given threat, through the memorized information about similar events in the past. For example, writing an angry text to its owner about the low running battery might result in a faster action of the owner, compared to a neutral or kind formulation. Regarding the software quality pillars defined in ISO/IEC 25010,[412] such individualized fear responses will likely lead to an increased software quality of the entire AI system, as it will choose actions to minimize any unpleasant quality experiences.

Another limitation of this fear circuit is that the saved time of how long an execution lasted, is not used any further. This information is valuable, as it can be used to evaluate the effectiveness of different fear responses in similar situations. By using the reinforcement learning algorithm proposed in the previous argument, this loss of valuable information can be circumvented. As running for a longer time will result in more possibilities for AI systems to get a higher reward, the run time of an execution would indicate to the reinforcement learning algorithm, which response is best suited in each situation.

In short, all limitations caused by the manual and sometimes arbitrary remapping from the biological onto the computational environment can be solved using dedicated hardware, reinforcement learning and authorization or user roles. In the case of reinforcement learning, this approach could help find an optimal solution for certain thresholds for a particular hardware and use case in an automated manner.

5.2 Neurological implications

The following limitations are all caused by uncertainties in the neurological basis, which are used as foundations for the creation of the fear circuit code.

First, neither learned nor innate fear are fully understood yet. State-of-the-art models of both types of fear build upon some assumptions, which remain to be verified. As neurologists are more focused on understanding every detail about learned fear than innate fear, the amount of research between both differs significantly.[413] Thus, it is likely that some details about the brain circuits of learned and especially innate fear might be missing in the fear circuit code. As both circuits are only implemented on a high-level of abstraction, which builds upon observations tested throughout hundreds of years,[414] new insights will likely not contradict the overall functionality of the fear circuit. Still, new research on biological fear circuits should be watched closely.

---

[412] Cf. ibid.
[413] Cf. *Rosén, Jörgen,* Fear, 2019, p. 13.
[414] Compare *Hoehl, S. et al.*, Innate Fear, 2017, p. 136; *Watson, J., Rayner, R.,* Little Albert, 1920, p. 10-12.

Similarly, the underlying brain circuits, which are not fully understood on their own,[415] are constructed on observations made in rodents.[416] Once more research about the human fear circuits will be available, slight modifications to the fear circuit code might need to be made. This is because the human fear circuit, as any other brain circuit in the human brain, is likely more complex and efficient than in other animals. Still, current research indicates that the high-level abstraction of the fear circuits used in this research holds true for humans as for rodents and other animals.[417]

Despite all these important limitations, this research provides important insights to further advance Artificial Intelligences and thus understand their implications on humanity. The goal of this research is not to provide a fully functionable AI with all sorts of emotions, but to provide a solid foundation for further research to advance AI. Although it is likely, that the main phenomena implemented into the fear circuit will not be affected by new neuroscientific research on lower-level details of the biological fear circuit, new finding might necessitate adaptions.

[415] Cf. *Mattera, A. et al.,* Learned Fear, 2020, p. 2.
[416] Cf. *Rosén, Jörgen,* Fear, 2019, p. 14-21.
[417] Cf. *Silva, B. et al.,* Innate Fear, 2016, p. 552.

6 Conclusion

Artificial Intelligence is a field of computer science that aims to create human-level intelligence in artificial agents.[418] Although many applications of AI outperform humans at specific tasks like playing chess,[419] they are not yet able to transfer their knowledge onto other tasks like humans.[420] They are prone to mistakes that would likely never happen to humans,[421] and they need significant amounts of data to learn something.[422] Leading industry experts argue that advances in popular paradigms used in AI like Deep Learning are not sufficient to achieve the goal of AI.[423] Instead, they argue that the implementation of brain circuitries seems promising to counteract the limitations of current AI systems.[424] To investigate whether and how brain circuits can be implemented into an AI, this research integrated a fear circuit into a state-of-the-art AI model. The emotion fear and the specific AI were chosen purposefully. Fear is critical to human survival[425] and can influence a multitude of other brain circuitries.[426] This is because all conditions associated with fear involve sensory input, a threat evaluation, fear memory and behavioral responses.[427] The fear circuit code was implemented into a publicly available AI chatbot with an open-source license.[428] The chatbot is based on a state-of-the-art transformer architecture from Microsoft Research.[429] To create the fear circuit code, elements of the biological fear circuit had to be remapped onto a computational environment. This included the specific stimuli, memorization unit, fear responses and the requirements for eliciting these responses. The symphony of the neurological processes involved in fear and the results of this remapping were composed into a logical model, which was implemented into code. This code was integrated with the chatbot's code. Thus, this research proved that it is possible to implement emotions, specifically fear, into an AI system.

---

[418] Cf. *Armstrong, S.,* Artificial Intelligence, 2017, p. 1.
[419] Cf. *Kurzweil, R.,* Artificial Intelligence, 2006, p. 2.
[420] Cf. *Kurzweil, R.,* Artificial Intelligence, 2006, p. 3.
[421] Cf. *Su, J. et al.,* One Pixel, 2019, p. 828-829.
[422] Cf. *Floridi, L., Chiriatti, M.,* GPT-3, 2020, p. 684.
[423] Cf. *Heath, N.,* Hassabis Interview, 2018, no page number.
[424] Cf. ibid.
[425] Cf. *LeDoux, J.,* Emotion, 2012, p. 660.
[426] Cf. *Maren, S.,* Fear, 2001, p. 903.
[427] Compare *Silva, B. et al.,* Innate Fear, 2016, p. 545; *Mattera, A. et al.,* Learned Fear, 2020, p. 3.
[428] Cf. *GitHub*, Gpt2bot, 2021, no page number.
[429] Cf. *Zhang, Y. et al.,* DialoGPT, 2020, p. 271.

In chapter 1.2, the following research question was established: Is it possible to implement the emotion fear into Artificial Intelligence? Given the results of this thesis, it can be answered as follows: Yes, it is possible to implement fear into AI. Both requirements for achieving the goal of this study, to serve as a steppingstone for further research on implementing brain circuitry into AI, were fulfilled. First, a fear circuit code was created that accounts for all high-level phenomena and processes involved in the creation, expression and memorization of fear. Second, it was successfully implemented into program code.

Despite the need for a remapping of some elements from a biological onto a computational environment, key phenomena associated with the fear circuit are observable in their coded version exactly as in their original setting.

This thesis highlights many limitations of the taken approach of implementing fear into an AI. They can be subdivided into constraints caused by the manual remapping and by limited neuroscientific insight on lower-level processes involved with fear. None of these limitations are severe enough to hinder the implementation itself, but they impede creating more intelligent AI. Therefore, this study lists several solutions on how to address these limitations in further research. These solutions involve software and hardware modifications. On the software side, the use of reinforcement learning at some parts of the circuit would ensure the input and output of the circuit can be adjusted more effectively than with manual, human work. This could increase the overall software quality of the AI as well.[430] The use of authorization and user roles will help to set up a range of safe and effective fear response behaviors. On the hardware side, the use of dedicated hardware for each brain circuitry might be beneficial to ensure a seamless interplay. Additionally, it is important to closely monitor fear research and adjust in the fear circuit code as needed.

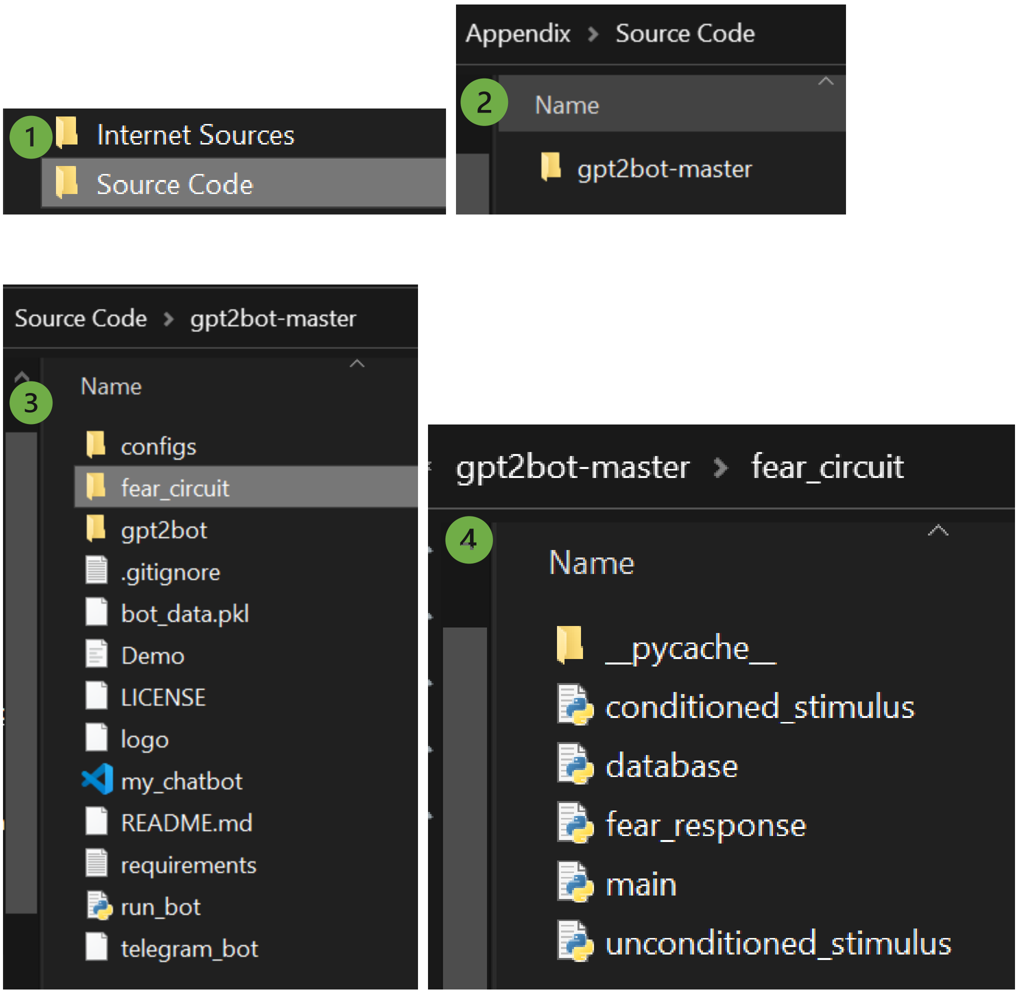
Further research should implement some of the proposed solutions and additional ones into the fear circuit code to ensure it is more suitable for the mass market. Ideally, these improvements should be publicly accessible to ease the implementation of further brain circuitries into an AI and to provide guidance. Once other brain circuits are integrated in AI systems, there will be a need for merging certain aspects of these circuits. For example, multiple circuits might need to access the same

[430] Cf. *Castillo, I. et al.,* Software Quality, 2010, p. 70.

sensory information or memorization system. Solutions must be established in advance on how multiple brain circuits can use the same resources. Additionally, a standardized guideline for the level of detail required for each brain circuit must be established. This will help to connect different brain circuits to each other more easily and with more neuroscientific profoundness.

Appendix

The source code is located in the attached folder “Source Code”> “gpt2bot-master”. It contains the fear circuit code and the Telegram chatbot. To access the fear circuit code itself, the folder “fear circuit” must be opened. The steps necessary to display the fear circuit code are shown in the following screenshots:



Bibliography

Armony, Jorge, L. (Emotion, 2005): Computational Models of Emotion, in: IJCNN, 3 (2005), p.1598-1602

Armstrong, Stuart (Artificial Intelligence, 2017): Introduction to the Technological Singularity, in: *Callaghan, Victor, Miller, James, Yampolskiy, Roman, Armstrong, Stuart* (Eds.) (Technological Singularity, 2017): The Technological Singularity Managing the Journey, 2017, p. 1-8

Barrett, Lisa F., Mesquita, Batja, Ochsner, Kevin N., Gross, James J. (Emotion, 2007): The Experience of Emotion, in: Annu Rev Psychol., 58 (2007), p. 373-403

Berlyne, D. E. (Curiosity, 1954): A theory of human curiosity, in: Brit Jour Psych, 45 (1954), no. 3, p. 180-191

Bhasin, H. (Python, 2019): Python Basics: A Self-Teaching Introduction, Dulles: Mercury Learning & Information, 2019

Callaghan, Victor, Miller, James, Yampolskiy, Roman, Armstrong, Stuart (Eds.) (Technological Singularity, 2017): The Technological Singularity Managing the Journey, Berlin: Springer, 2017

Calvo, Rafael, D'Mello, Sidney, Gratch, Jonathan, Kappas, Arvid (Eds.) (Affective Computing, 2015): The Oxford Handbook of Affective Computing, New York: Oxford University Press, 2015

Castillo, Isi, Losavio, Francisca, Matteo, Alfredo, Bøegh, Jørgen (Software Quality, 2010): REquirements, Aspects and Software Quality: the REASQ model, in: Jour O T, 9 (2010), no. 4, p. 69-91

Dalgleish, Tim, Power, Mick (Eds.) (Emotion, 1999): Handbook of Cognition and Emotion, New York: John Wiley & Sons, 1999

Darwin, Charles (Emotion, 1872): The Expression of the Emotions in Man and Animals, London: John Murray, 1872

Davis, Michael (Fear, 1992): The role of the Amygdala in fear and anxiety, in: ARN, 15 (1992), p. 353-375

Dolcos, Florin, Katsumi, Yuta, Weymar, Mathias, Moore, Matthew, Tsukiura, Takashi, Dolcos, Sanda (Fear, 2017): Emerging Directions in Emotional Episodic Memory, in: Fr Psych, 8 (2017), no.1867, p. 1-25

Du, Shichuan, Martinez, Aleix M. (Emotion, 2015): Compound facial expressions of emotion: from basic research to clinical applications, in: Dia C Neuro, 17 (2015), no. 4, p. 443-455

*Ekman, Paul (*Darwin, 2009): Darwin's contributions to our understanding of emotional expressions, in: Phil TRS B, 364 (2009), p. 3449-3451

Ekman, Paul (Emotion, 1972): Universals and Cultural Differences in Facial Expressions of Emotions, in: Neb S M, 19 (1972), p. 207-282

Ekman, Paul (Emotion, 1992): An argument for basic emotions, in: Cog Emot, 6 (1992), p. 169-200

Ekman, Paul (Emotion, 1999): Basic emotions, in: *Dalgleish, Tim, Power, Mick* (Eds.), Handbook of Cognition and Emotion, 1999, p. 45-60

Erickson, Robert P. (Intelligence, 2014): Are Humans the Most Intelligent Species?, in: J Intell, 2 (2014), p.119-121

Ferry, Barbara (Ed.) (Amygdala, 2017): The Amygdala - Where Emotions Shape Perception, Learning and Memories, Rijeka: InTech, 2017

Floridi, Luciano, Chiriatti, Massimo (GPT-3, 2020): GPT-3: Its Nature, Scope, Limits, and Consequences, in: MiMa, 30 (2020), p. 681-694

Goleman, Daniel (Emotional Intelligence, 1996): Emotional Intelligence, New York: Bantam, 1996

Hadzhieva, Tsvetelina S. (Emotion, 2017): Conscious and Unconscious Emotions in Alexithymics and Repressors, in: Psych Tho, 10 (2017), no. 1, p. 227-235

Helm, Bennett W. (Emotion, 2009): Emotions as Evaluative Feelings, in: Em Rev, 1 (2009), p. 248-255

*Hoehl, Stefanie, Hellmer, Kahl, Johansson, Maria, Gredebäck, Gustaf (*Innate Fear, 2017): Itsy Bitsy Spider…: Infants React with Increased Arousal to Spiders and Snakes, in: Fr Psych, 8 (2017), no. 1710, p. 130-138

Huber, Daniel, Veinante, Pierre, Stoop, Ron (CeM, 2005): Vasopressin and Oxytocin Excite Distinct Neuronal Populations in the Central Amygdala, in: SCI, 308 (2005), no. 8, p. 245-248

Izquierdo, Ivan, Furini, Cristiane R. G., Myskiw, Jociane C. (Fear Memory, 2016): Fear Memory, in: Phys Rev, 96 (2016), no. 2, p. 695-750

Kalliamvakou, Eirini, Gousios, Georgios, Blincoe, Kelly, Singer, Leif, German, Daniel M., Damian, Daniela (GitHub, 2014): The promises and perils of mining GitHub, in: ICSE, 36 (2014), p. 92-101

Kima, Jeansok J., Jungb, Min W., (Pavlovian fear, 2006): Neural circuits and mechanisms involved in Pavlovian fear conditioning: A critical review, in: NEU B REV, 30 (2006), no. 2, p. 188-202

Krizhevsky, Alex, Sutskever, Ilya, Hinton, Geoffrey E. (ImageNet, 2012): ImageNet Classification with Deep Convolutional Neural Networks, in: ANIPS, 25 (2012), p. 1106-1114

Kurzweil, Ray (Artificial Intelligence, 2006): The Singularity Is Near: When Humans Transcend Biology, New York: Penguin Books, 2006

LeDoux, Joseph E., (Emotion, 2012): Rethinking the emotional brain, in: NRN, 73 (2012), no. 3, p. 653-676

Li, Guoshi (Learned Fear, 2017): Computational Models of the Amygdala in Acquisition and Extinction of Conditioned Fear, in: *Ferry, Barbara* (Ed.), The Amygdala - Where Emotions Shape Perception, Learning and Memories, 2017, p. 235-258

Maren, Stephen (Fear, 2001): Neurobiology of Pavlovian fear conditioning, in: ARN, 24 (2001), p. 897-931

Maren, Stephen, Quirk, Gregory J. (Fear, 2004): Neuronal signalling of fear memory, in: N R NEURO, 5 (2004), no.11, p. 844-852

Martin, Robert C. (Clean Code, 2008): Clean Code A Handbook of Agile Software Craftsmanship, Boston: Pearson Education, 2008

Mattera, Andrea, Pagani, Marco, Baldassarre, Gianluca (Learned Fear, 2020): A Computational Model Integrating Multiple Phenomena on Cued Fear

Conditioning, Extinction, and Reinstatement, in: Fr Sys Neuro, 14 (2020), no. 569108, p. 1-22

Minsky, Marvin (Emotion, 2007): The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind, 2. Ed., New York: Simon & Schuster, 2007

Mobbs, Dean, Adolphs, Ralph, Fanselow, Michael S., Barrett, Lisa F., LeDoux, Joseph E., Ressler, Kerry, Tye, Kay M. (Fear, 2019): Viewpoints: Approaches to defining and investigating fear, in: N NEURO, 22 (2019), p. 1205–1216

Mohri, Mehryar, Rostamizadeh, Afshin, Talwalkar, Ameet (Machine Learning, 2018): Foundations of Machine Learning, 2. Ed., Cambridge, London: MIT Press, 2018

Picard, Rosalind W. (Affective Computing, 2015): The Promise of Affective Computing, in: *Calvo, Rafael, D'Mello, Sidney, Gratch, Jonathan, Kappas, Arvid* (Eds.), The Oxford Handbook of Affective Computing, 2015, p. 11-20

Reddi, Vijay J., Plancher, Brian, Kennedy, Susan, Moroney, Laurence, Warden, Pete, Agarwal, Anant Banbury, Colby, Banzi, Massimo, Bennett, Matthew, Brown, Benjamin, Chitlangia, Sharad, Ghosal, Radhika, Grafman, Sarah, Jaeger, Rupert, Krishnan, Srivatsan, Lam, Maximilian, Leiker, Daniel, Mann, Cara, Mazumder, Mark, Pajak, Dominic, Ramaprasad, Dhilan, Smith, J. Evan, Stewart, Matthew, Tingley, Dustin (TinyML, 2021): Widening Access to Applied Machine Learning with TinyML, in: CoRR, 2106.04008 (2021), p. 1-10

Rosén, Jörgen (Fear, 2019): Innate and Conditioned Fear Investigating Responses to Threat using Psychophysiology, Functional Magnetic Resonance Imaging, and Twin Methodology, Uppsala: Acta Universitatis Upsaliensis, 2019 (Dissertation, Univ. Uppsala 2019)

Russell, Stuart J. (Artificial Intelligence, 2020): Human Compatible: Artificial Intelligence and the Problem of Control, New York: Penguin, 2020

Russell, Stuart J., Norvig, Peter (Artificial Intelligence, 2021): Artificial Intelligence, A Modern Approach, 4. Ed., Harlow: Pearson Education, 2021

Shalev-Shwartz, Shai, Ben-David, Shai (Machine Learning, 2014): Understanding Machine Learning From Theory to Algorithms, New York: Cambridge University Press, 2014

Shin, Lisa M., Rauch, Scott L., Pitman, Roger K. (Amygdala, 2006): Amygdala, Medial Prefrontal Cortex, and Hippocampal Function in PTSD, in: ANYAS, 1071 (2006), no. 1, p. 67-79

Silva, Bianca A., Gross, Cornelius T., Gräff, Johannes (Innate Fear, 2016): The neural circuits of innate fear: detection, integration, action, and memorization, in: Learn Mem, 23 (2016), no. 10, p. 544-555

Stone, Peter, Brooks, Rodney, Brynjolfsson, Erik, Calo, Ryan, Etzioni, Oren, Hager, Greg, Hirschberg, Julia, Kalyanakrishnan, Shivaram, Kamar, Ece, Kraus, Sarit, Leyton-Brown, Kevin, Parkes, David, Press, William, Saxenian, AnnaLee, Shah, Julie, Tambe, Milind, Teller, Astro (AI Report, 2016): Artificial Intelligence and Life in 2030 One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford: Stanford University, 2016

Su, Jiawei, Vargas, Danilo V., Kouichi, Sakurai (One Pixel, 2019): One Pixel Attack for Fooling Deep Neural Networks, in: IEEE TEC, 23 (2019), no. 5, p. 828-841

Totterdell, Peter, Hershcovi, M. Sandy, Niven, Karen, Reich, Tara C., Stride, Chris (Emotional drain, 2012): Can employees be emotionally drained by witnessing unpleasant interactions between co-workers? A diary study of induced emotion regulation, in: Wo Str, 26 (2012), no. 2, p. 112-129

Turing, Alan M. (Computing, 1950): Computing Machinery and Intelligence, in: MND, 59 (1950), no. 236, p. 433-460

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia (Transformer, 2017): Attention Is All You Need, in: ANIPS, 30 (2017), p. 5998-6008

Vertes, Robert P. (IL PL, 2004): Differential Projections of the Infralimbic and Prelimbic Cortex in the Rat, in: SYN, 51 (2004), no. 1, p. 32-58

Watson, John B., Rayner, Rosalie (Little Albert, 1920): Conditioned emotional reactions, in: Jour E P, 3 (1920), no. 1, p. 1-14

Zadra, Jonathan R., Clore, Gerald L. (Emotion, 2011): Emotion and Perception: The Role of Affective Information, in: W IRC Sci, 2 (2011), no. 6, p. 676-685

Zhang, Yizhe, Sun, Siqi, Galley, Michel, Chen, Yen-Chun, Brockett, Chris, Gao, Xiang, Gao, Jianfeng Liu, Jingjing, Dolan, Bill (DialoGPT, 2020): DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation, in: ACL, 58 (2020), p. 270-278

Internet sources

Battery University (Battery, 2019): BU-808: How to Prolong Lithium-based Batteries (2019-07-10), https://batteryuniversity.com/article/bu-808-how-to-prolong-lithium-based-batteries, (Access 2021-08-28, 09:33 MEZ)

Brin, Sergey (Google, 2017): 2017 Founders' Letter (n.d.), https://abc.xyz/investor/founders-letters/2017/index.html, (Access 2021-07-29, 11:03 MEZ)

Cellan-Jones, Rory (Hawking, 2014): Stephen Hawking warns artificial intelligence could end mankind (2014-12-02), https://www.bbc.com/news/technology-30290540, (Access 2021-07-29, 11:51 MEZ)

Ekman, Paul (Emotion Duration, n.d.): Mood vs. Emotion: Differences & Traits (n.d.), https://www.paulekman.com/blog/mood-vs-emotion-difference-between-mood-emotion/, (Access 2021-07-01, 13:57 MEZ)

Ekmann, Paul (Emotion, n.d.): What are emotions? (n.d.), https://www.paulekman.com/universal-emotions/, (Access 2021-07-01, 13:40 MEZ)

European Commission (Innate Fear, 2016): Functional wiring of the core neural network of innate fear (2016-05-18), https://cordis.europa.eu/project/id/341139, (Access 2021-09-09, 10:44 MEZ)

Forbes (Python, 2020): Python Language: What You Need To Know (2020-07-24), https://www.forbes.com/sites/tomtaulli/2020/07/24/python-language-what-you-need-to-know/?sh=6aeeb639a922, Access (2021-08-09, 19:24 MEZ)

Frontiers (Emotions, 2018): How Many Different Kinds of Emotion are There? (2018-05-09), https://kids.frontiersin.org/articles/10.3389/frym.2018.00015, (Access 2021-07-01, 14:12 MEZ)

GeeksforGeeks (self Python, 2021): self in Python class (2021-07-05), https://www.geeksforgeeks.org/self-in-python-class/, (Access 2021-10-21, 20:33 MEZ)

GitHub (Gpt2bot, 2021): Gpt2bot (2021-01-25), https://github.com/polakowo/gpt2bot, (Access 2021-10-13, 20:19 MEZ)

GitHub (MIT License, 2021): MIT-LICENSE.txt (2021-01-03), https://github.com/git/git-scm.com/blob/main/MIT-LICENSE.txt, (Access 2021-10-15, 17:37 MEZ)

Heath, Nick (Hassabis Interview, 2018): Google DeepMind founder Demis Hassabis: Three truths about AI (2018-09-24), https://www.techrepublic.com/article/google-deepmind-founder-demis-hassabis-three-truths-about-ai/, (Access 2021-09-24, 08:55 MEZ)

High, Peter (Artificial Intelligence, 2017): Carnegie Mellon Dean Of Computer Science On The Future Of AI (2017-10-30), https://www.forbes.com/sites/peterhigh/2017/10/30/carnegie-mellon-dean-of-computer-science-on-the-future-of-ai/?sh=d0b2bf221979, (Access 2021-08-10, 20:29 MEZ)

HP (HP Battery, 2020): How Long Should I Charge my HP Laptop Battery? (2020-10-28), https://www.hp.com/in-en/shop/tech-takes/post/how-long-should-i-charge-hp-laptop-battery, (Access 2021-08-28, 09:17 MEZ)

Intel (CPU, n.d.): How to Fix High CPU Usage (n.d.), https://www.intel.com/content/www/us/en/gaming/resources/how-to-fix-high-cpu-usage.html, (Access 2021-09-27, 15:11 MEZ)

Intel (High CPU, 2021): What Are the Consequences of High Processor Usage for Long Periods of Time? (2021-07-23), https://www.intel.com/content/www/us/en/support/articles/000038081/processors.html, (Access 2021-09-27, 15:24 MEZ)

Max Planck Institute (Innate Fear, 2017): Itsy Bitsy Spider: Fear of spiders and snakes is deeply embedded in us (2017-10-19), https://www.cbs.mpg.de/Fear-of-spiders-and-snakes-is-deeply-embedded-in-us, (Access 2021-08-28, 13:12 MEZ)

McEachron, Coralie (Four F's, 2021): Why We Do What We Do: Trauma and Stress Responses – The Four F’s (2021-04-26), http://northstarcounselingcenter.com/trauma-and-stress-responses-the-four-4s, (Access 2021-07-01, 14:24 MEZ)

MySQL (MySQL, n.d.): About MySQL (n.d.), https://www.mysql.com/about/, (Access 2021-10-20, 08:48 MEZ)

New York Times (Emotional Behavior, 1993): 'Miracle Child' Is Survivor (1993-09-25), https://www.nytimes.com/1993/09/25/us/miracle-child-is-survivor.html, (Access 2021-07-26, 11:09 MEZ)

Peck, Danni (Learned Fear, 2021): Conditioned Stimulus And Psychology (2021-09-20), https://www.betterhelp.com/advice/general/conditioned-stimulus-definition-and-its-role-in-psychology/, (Access 2021-09-04, 07:45 MEZ)

Psutil (Package psutil, n.d.): psutil documentation (n.d.), https://psutil.readthedocs.io/en/latest/, (Access 2021-20-21, 20:21 MEZ)

PyPi (Package screen_brightness_control, 2021): screen_brightness_control (2021-10-30), https://pypi.org/project/screen-brightness-control/, (Access 2021-10-21, 21:07 MEZ)

Python (Package time, 2021): time — Time access and conversions (2021-11-01), https://docs.python.org/3/library/time.html, (Access 2021-10-17, 13:55 MEZ)

RT.com (Putin, 2017): 'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day (2017-09-01), https://www.rt.com/news/401731-ai-rule-world-putin/, (Access 2021-07-29, 11:18 MEZ)

Scott, Kevin (GPT-3, 2020): Microsoft teams up with OpenAI to exclusively license GPT-3 language model (2020-09-22), https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/, (Access 2021-08-12, 12:31 MEZ)

Shead, Sam (Elon Musk, 2020): Elon Musk has a complex relationship with the A.I. Community (2020-05-13), https://www.cnbc.com/2020/05/13/elon-musk-has-a-complex-relationship-with-the-ai-community.html, (Access 2021-07-29, 17:48 MEZ)

TechVidvan (Python Methods, n.d.): Python Methods vs Functions – What really differentiates them? (n.d.), https://techvidvan.com/tutorials/python-methods-vs-functions/, (Access 2021-08-12, 16:33 MEZ)

Telegram (Telegram Bot, n.d.): Bots: An introduction for developers (n.d.), https://core.telegram.org/bots, (Access 2021-10-13, 14:25 MEZ)

Tesla (Elon Musk, n.d.): Elon Musk (n.d.), https://www.tesla.com/elon-musk, (Access 2021-07-12, 08:34 MEZ)

Wikimedia Commons (Classical Conditioning, 2016): File: Classical Conditioning Diagram.png (2016-11-23), https://commons.wikimedia.org/w/index.php?curid=53458610, (Access 2021-11-01, 08:53 MEZ)

Woodfine, Gary (Clean Code, 2018): Book Review: Clean Code (2018-10-18), https://garywoodfine.com/book-review-clean-code/, Access (2021-08-09, 17:21 MEZ)

Declaration in lieu of oath

I hereby declare that I produced the submitted paper with no assistance from any other party and without the use of any unauthorized aids and, in particular, that I have marked as quotations all passages, which are reproduced verbatim or near-verbatim from publications. Also, I declare that the submitted print version of this thesis is identical with its digital version. Further, I declare that this thesis has never been sub-mitted before to any examination board in either its present form or in any other similar version. I herewith agree that this thesis may be published. I herewith consent that this thesis may be uploaded to the server of external contractors for the purpose of submitting it to the contractors' plagiarism detection systems. Up-loading this thesis for the purpose of submitting it to plagiarism detection systems is not a form of publication.

Stuttgart, 04.11.2021