

FUNCTIONALIST EMOTION MODEL IN ARTIFICIAL GENERAL INTELLIGENCE

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Xiang Li
August, 2021

Examining Committee Members:

Hongchang Gao, Advisory Chair, Dept. of Computer and Information Sciences, Temple University

Pei Wang, Advisor, Dept. of Computer and Information Sciences, Temple University

Hongling Xie, Dept. of Psychology, Temple University

Antonio Chella, Dep. of Engineering, University of Palermo

For all the people

ABSTRACT

Functionalist Emotion Model in Artificial General Intelligence

by

Xiang Li

The objective of this research is to elucidate motivation and emotion processing in an AGI (Artificial General Intelligence) system NARS (Non-Axiomatic Reasoning System). Under the basic assumption that an artificial general intelligence system should work with insufficient resources and knowledge, the emotion module can help direct the selection of internal tasks, and allow the autonomous allocation of internal resources and rapid response with urgency, so that the inference capability of AGI system can be improved.

The psychological and AI theories related to emotion are extensively reviewed, including the source of emotion, the appraisal process in emotional experience, the cognitive processing and coping process, and the necessity of emotion for Artificial General Intelligence design.

This dissertation describes the conceptual design, realization process and application process of emotion in NARS. The process of internal resource allocation triggered by different emotions based on NARS reasoning framework is proposed, and the design can be applied to any scene. The similarity and difference between human emotion and artificial intelligence emotion are discussed. At the same time, the advantages and disadvantages of the design and its theory are also discussed.

A recent implementation of the NARS model will be discussed with examples. and the emotion model has been tested preliminarily in a new version of OpenNARS. New *Temporal Induction* model, *Anticipation* model, *Goal* processing model, and *Emotion* models which are implemented in the new system will also be discussed in detail.

The dissertation concludes with suggestions and ideas that are put forward for the role of emotion in future human-computer interaction.

ACKNOWLEDGEMENTS

Before I really started working on artificial intelligence, my interest was always in developing video games. As someone who was introduced to the arcade by my cousin when I was three years old, video games can be said to have influenced my life from the time I started playing them to today. Right before I graduated from Bloomsburg University, I still wanted to continue my education in the field of video games in graduate school. But, at the time, as someone who had been playing video games for 20 years, I had witnessed a major revolution in video games. I still remember when I was 10 years old, I was completely immersed in the complex puzzle solving process of Resident Evil II. However, starting with Resident Evil IV, the improved graphics and feeling of involvement resulted in fewer puzzles and less complexity in the game. The best Japanese game designers had to compromise with the European and American game markets and start making games that were more suitable for American and European gamers. Since then, my thinking about getting into the field of video games has changed. The change in style of today's popular games seems to confirm my belief that, at the very least, I won't regret moving into a more challenging research field – Artificial General Intelligence.

I still remember when Dr. Pei Wang asked me whether my future research plan would be more theoretical or more practical. I seem to answer without hesitation, “practical”. I also remember that I chose artificial intelligence emotion without hesitation after Dr. Wang listed possible research directions. But after years of study in this field, I realized that compared with practical research, the study combining theory and practice is more difficult, but also more interesting.

Having graduated from Bloomsburg University and had started at Temple University, I was confident that having studied in the United States for three years, I

had advantages over some freshmen from China that they cannot have. However, the midterm exams in the first semester made me pay the price for my arrogance. At that time, I wanted to give up, but Dr.Wang pulled me back from the brink of despair.

I am very grateful to Dr.Wang for his guidance, support, and encouragement in the past seven years. At the same time, I also want to thank him for never giving up on teaching me. I am confident that I could be an excellent person in any field that I am interested, but I have to admit that I am not an excellent student of Dr. Wang. But, as always, Dr. Wang helped me like a friend and even a father.

As I said to Dr.Wang, I admired his knowledge of all areas of artificial intelligence, from computer science to philosophy and psychology. The complete artificial intelligence system in his heart has opened a new chapter in the field of artificial intelligence. I am very honored to be his student, and I hope that one day, I can have the same understanding and insight of artificial intelligence as him.

Special thanks to Dr.Hongling Xie for her support on emotional psychology in this project. I am also grateful to Dr.Hongchang Gao and Dr.Antonio Chella for their support and help, especially when I'm in trouble. Finally, I would like to thank Dr.Kris Thorrison, even though he cannot be the external reader of my dissertation. I still remember the day I discussed emotion of artificial intelligence with Kris in Czech Republic, which has played an important inspiration role in today's research.

I would also like to dedicate this dissertation to God, as a guide to my life every day in the past, and a light to my life in the future. I would also dedicate this paper to my family, my parents, Li Yanli and Huo Shaojing, for their help and encouragement in my education all the time. My future parents-in-law, I hope this paper will at least partially make them satisfied. My friend Patrick Hammer, who has patiently answered every single question I had with OpenNARS code issues, even though it was almost silly to him at the beginning. My friends, Meng Fanhua and Song Lirong, who provided help and support for my life in the last semester of pursuing my Ph.D.

I am very grateful to Christian Hahm for his help with the proofreading of this dissertation. Special thanks to my best friend in the United States, Lenny Oddo, for checking English grammar in this dissertation, and for the care and help that Lenny and his family have given me in the past ten years.

Finally, I would like to thank my fiancée, Liu Meijia, who has been enduring my anxiety brought by the pressure of life and study. She has always been praying for me, encouraging me and helping me. She contributed and sacrificed a lot, and I hope I didn't waste or disappoint her love for me.

TABLE OF CONTENTS

DEDICATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENTS	v
LIST OF FIGURES	x
LIST OF TABLES	xi
CHAPTER	
1. INTRODUCTION	1
2. EMOTION IN PSYCHOLOGY	5
2.1 Human Scenario	5
2.2 What is Emotion?	7
2.3 Which Theory is Suitable for AI emotion?	10
3. EMOTION IN AI	18
3.1 Why Do AI Scholars “Shy Away” From Emotions?	18
3.2 Why do we need Emotion in AGI	19
3.3 Previous Works on Emotion in AGI	20
4. NON-AXIOMATIC REASONING SYSTEM	30
4.1 NARS Overview	30
4.2 Inference control	34
4.3 New Architecture	35
5. REQUIREMENTS FOR AN AGI EMOTION MECHANISM	41
5.1 Design Requirements	42
5.2 Functional Requirements	47
5.3 Architecture Requirements	50
5.4 Summary	54

6. CONCEPTUAL DESIGN OF EMOTION MODEL IN NARS	55
6.1 Should Emotion be Innate or Learned in NARS?	55
6.2 Architecture of Emotion Module in NARS	59
6.3 How many locks can a “key” open?	64
6.4 The Functions of Emotion	73
6.5 Summary	75
7. EMOTION MODEL IN NARS	77
7.1 Structure of Buffers	77
7.2 Anticipation	85
7.3 Aircraft Combat Game as Testing Case	87
7.4 Emotion Mechanism in NARS	90
7.5 Summary	103
8. COMPARISON WITH EXISTING ARCHITECTURES	106
8.1 Affective Computing	106
8.2 ACT-R	107
8.3 CLARION	108
8.4 LIDA	110
8.5 MicroPsi	111
8.6 Sigma	112
8.7 NARS	112
8.8 Summary	113
9. CONCLUSION AND FUTURE WORKS	115
9.1 Major Results	115
9.2 Limitations and Future Works	117
BIBLIOGRAPHY	120
APPENDIX	126

LIST OF FIGURES

Figure

4.1	The old Architecture of Control Mechanism in NARS	36
4.2	The New Architecture of Control Mechanism in NARS	39
6.1	Task flow and Emotional flow within the target framework	62
7.1	Example state of a buffer	78
7.2	Temporal Induction in the Event Sequence of Buffer	82
7.3	The aircraft combat game	88
7.4	Rank of Learned Knowledge at Early State of the Test Example . .	95
7.5	Rank of Learned Knowledge at Late State of The Test Example . .	96
7.6	Testing Score Without Emotion After 200 seconds	98
7.7	Testing Score With Emotion After 200 seconds	99
7.8	Internal view of the Buffer	105

LIST OF TABLES

Table

6.1	The relation among desire-value, truth-value, and the valence	66
6.2	XNOR Gate Truth Table	66
6.3	Integrative model of Happiness, Sadness, Anger and Fear	68

CHAPTER 1

INTRODUCTION

No one has made a good argument against the importance of emotions in human life, but the AI community has always kept emotions at arm’s length. It is often said that the presence of emotion is the most important difference between a human and a machine, but this idea may directly stifle the possibility of emotion being realized in artificial intelligence systems. Even the designers of OCC cognitive models (*Ortony et al.*, 1988) of emotion, after implementing their emotion appraisal model with computers, still consider it pointless and futile to achieve true emotion through artificial intelligence or create an agent with emotion (*Ortony et al.*, 1988).

Emotional expression is the most obvious of the many emotional functions, it plays a very important role in the adaptability of the individual, but it is not the only function of emotion. Other important functions of emotion include resource allocation, motivation generation, fast reaction etc. These functions require individuals to complete emotional appraisal at the abstract level — in other words, as long as the event meets the criteria of certain emotional appraisal models, the corresponding emotion can be triggered, and allow the emotion to make changes to the internal environment of the individual. Such a mechanism does not require the system to learn an “event-emotion-coping” mapping, only to know the appraisal conditions of different emotions such that the system can appraise any incoming events.

The research in this dissertation is an attempt to explore how emotion affects the inference capability of artificial general intelligence, and its purpose is to break the long-term attitude of artificial intelligence system to “keep away from emotion”, to

prove the necessity of emotion in artificial general intelligence, as well as the possibility of realization.

In order to implement emotional functions in artificial general intelligence, research needs to be carried out at three levels of abstraction.

First, many philosophical, psychological, and methodological questions need to be considered. What is emotion? What is the role of emotion in human beings? What are the advantages and disadvantages of having emotions? Is it possible to have emotional functionality in an AI system? Is it necessary to have emotional functionality in an AI system? Why would AI systems have emotions? How to evaluate an artificial intelligent system to have emotion? How should emotions help artificial intelligent systems? How to control artificial intelligence system not to abuse emotion? Though there is still no consensus on these questions in the field, my views on these issues set the foundation for the entire research project by providing clear evidence from different disciplinary perspectives.

Second, the human emotion model is used as the theoretical basis and reference frame for the implementation of artificial intelligence emotion model. The human emotions have been studied for hundreds of years. Although philosophers, psychologists, and neuroscientists have not fully explored all the aspects about emotions, the existing research can provide a sufficient theoretical framework for our research. Although this study uses human emotions as a reference, but the purpose is not to copy human emotion model, or simulate exactly the same emotion process in the artificial intelligence system, the fundamental reason is that there are insurmountable gaps between human and machine, the evaluation of a same condition might be completely different from the view of artificial intelligent agent and human on the same thing, even said, artificial intelligence agents may generate independent emotional systems over a long period of learning and adaptation. Therefore, this study aims to take the framework of human emotion as a reference and achieve functional emotion

as the purpose. This project mainly studies how emotions would help the artificial general intelligent systems to improve the autonomy, as well as improve the ability of learning, decision-making and reasoning.

Finally, the entire model needs to be implemented in a computer system. All described models should be embedded in programs, which must then be run in the available software and hardware environments.

The organization of this dissertation is based on these three levels.

Chapter 2 introduces the theoretical foundation of the project. Through the analysis of three important branches of emotion research in psychology, it is concluded that the study of the Cognitive Theory branch represented by Lazarus is more suitable as the theoretical foundation of artificial general intelligence emotion research.

Chapter 3 briefly describes the current AI community's attitude towards emotion and the reasons for keeping emotion at a distance, and briefly expounds why emotion is necessary for general AI systems. Also briefly introduces related works on emotion study in current artificial intelligence field and artificial general intelligence field.

Chapter 4 mainly introduces Non-Axiomatic Logic (NAL) and its corresponding implementation, Non-Axiomatic Reasoning System (NARS). This chapter does not include detailed introduction about Logic and Control mechanism, but only contains part of the content, aiming to help readers understand the following contents of this paper. For a detailed description of NAL and NARS, please read (*Wang, 2006*), and (*Wang, 2013*) and the related studies on GitHub.

Chapter 5 describes the requirements for emotion modules to be added to the artificial general intelligent system. This chapter starts from three aspects, namely, **Design requirements**, elaborates the relationship between emotion modules and the main body of the intelligent system. **Functional requirements**, introduce the function the AI emotion module should have, or how emotion should help the intelligent system; The **Architecture requirements**, introduce what the architecture

of an emotion model in artificial general intelligent system should look like. All the discussions in this chapter are based on the summary of the NARS emotion module design and the requirements reflected in the related work, and are not aimed at all intelligent systems.

Chapter 6 introduces the conceptual design of emotion module in NARS. This chapter has the construction of emotion module, the realization process, and the function of emotion module and the discussion of related issues.

Chapter 7 describes the specific implementation of emotion module in NARS. A new control architecture supporting emotion module will be introduced in detail in this chapter. Readers can have a detailed understanding of the new control architecture of NARS, how emotion module is implemented in NARS, and the main functions of emotion module in NARS. Several experimental results are shown along the description of the new architecture, which proves the necessity of emotion in NARS.

Chapter 8 will provide a brief comparison between the emotion module of NARS and the related work described in Chapter 3.

Finally, Chapter 9 will summarize the project, give the outlook future research work of artificial general intelligence emotion, and elaborate on the related work in the future.

CHAPTER 2

EMOTION IN PSYCHOLOGY

2.1 Human Scenario

In the following circumstance, consider the situation in which a paladin, Troy, is in danger while escorting Doctor Tom back to the castle, who can save the princess. The princess was ill, and only the magic potion of Dr. Tom in another town could cure her. Tom's town was three days' ride from the princess's. Troy had been walking with Tom for two days, and if all went well, they would both be back in the castle by noon tomorrow. It was now the night of the second day, and because they were eager to travel in the daytime, they had not had much to eat or rest, and both the horses and the men were very hungry and tired. Troy and Tom decided to light a fire in the jungle to rest. Tom took out the meal that his wife prepared for him and Troy to replenish their energy. The two men sat down on the ground and had barely taken a bite when they heard a strange noise in the bushes near them. As a knight, Troy instinctively drew his sword, stood up to meet the possible danger, and Tom also clenched the bread in his hand and moved carefully behind Troy. The sound died away, and Troy thought it had just been a squirrel. They went back to the fire. Hungrier than ever after their fright, they took out their food again and began to eat. Suddenly, a dark figure rushed out of the bushes and jumped towards Troy. Troy immediately threw away the food and avoided the shadow. After avoiding the shadow, Troy immediately drew his sword to face the shadow. At this time, he saw clearly that the shadow was a huge wolf. Tom now stood in place and did not dare to move, while Troy had two choices at this time, to avoid the giant wolf by riding away,

and coming back in the morning to get the medicine left by Tom. But such a choice, is bound to give up Tom, because Tom in the face of the giant wolf, would surely result in death. The other option is to fight the giant wolf, while protecting Tom at the same time. Troy is not completely sure about this battle, the worst outcome is that they both die, and the potion can not be delivered to the princess in time. Troy decided to die to protect Tom since Tom saved his life in the battlefield. During a fierce battle, Troy and Tom worked together to kill the wolf. Troy got wounded, but when he sees that Doctor Tom was safe and had started treating him, Troy smiled peacefully.

There are a lot of emotional experiences in this short story. First of all, Troy and Tom were eager to arrive at the castle in time so they choose to shorten the time of rest and eat in the daytime. To some readers, this is kind of optimization process of decision making, the emotions play a key role in prioritizing the tasks. If they are late, and the princess doesn't get the drug in time, she will die, there will be a punishment from the king, Troy and Tom will have extreme sense of inner guilt. Their emotions could reduce their feeling of hunger and their feeling of fatigue (a desire to eat and rest). Emotions then allow them to increase their time spent walking during the day and decrease their resting time after sunset.

When Troy and Tom heard the sound in the bush for the first time, they were surprised and accompanied by fear. The different performances of the two in the face of fear, prove that different people will trigger different degrees of emotion and cope in different ways. Troy, who is experienced in hundreds of battles, chooses to face the fear. While Tom, who has no combat experience, chooses to stand behind Troy and seek shelter. At this time, they may have ignored their hunger and fatigue. The fear of losing their own life is heightened, which make them temporarily give up eating and resting. In addition to turning their attention to the thing that threatened their safety, and turning their attention to people who can help them to get rid of

the threat. Their different coping styles are based on their past experiences, with the battle-hardened Troy fighting when faced with danger, and healer Tom seeking shelter.

When the sound subsided temporarily, Troy and Tom felt a temporary recovery from their nervousness. The threat to the survival of the object disappeared. The fear then gradually subsided, and, eating and resting took precedence over defense and avoidance. However, when the wolf attack again, it stopped the desire of eating and resting again. Now, Troy needed to take the action avoid, and thanks to Troy's usual training and experience on the battlefield, Troy can stop pursuing the current goal of fulfilling himself by food and perceiving risk of being hurt in a very short time and to respond in a timely manner. The allocation of resources activated by the fear within the body, makes Troy's attention transfer to the object which threaten his life.

The fundamental reason for Troy to stay and fight with the Wolf and protect Tom is Troy's gratitude to Tom. Leaving would undoubtedly be the most rational choice for Troy, but Troy did not make the choice in accordance with rationality. Although Troy was injured in the end, the joy of winning the battle and successfully protecting Tom overshadowed the sadness activated by Troy's injury.

2.2 What is Emotion?

The focus of this paper is on the theoretical and practical research of artificial intelligent emotion. Before AI's work on emotions begins, the only source of emotion research is psychological research on emotion. Although with the continuous development of neuroscience in recent years, many neuroscientists also participated in the study of emotion, but the research is focused on when emotions are triggered by the corresponding brain structures, as well as the related potential changes (*Deak*, 2011; *Panksepp*, 2010). However, subjective experience of emotion, cognitive appraisal of different emotions, and behavior after different emotions are triggered, are mostly

come from psychology and cognitive psychology research.

To talk about emotion, we first need to know what emotion is, or how emotion is defined. According to Plutchik's statistics, there have been at least 90 different definitions of emotion in psychology throughout the history of emotion research (*Plutchik, 1982*).

This dissertation has no intention to discuss the validity, or the advantages and disadvantages of different definitions of emotion, because the structure of the human body and the structure of the machine are different. To evaluate the definition of human emotion from the perspective of artificial intelligence is indeed an unfair thing. Different researchers attach different emphasis to the definition of emotion according to different research directions. Some researchers may focus on subjective experiences and feelings, while others focus on neural activity and behavioral responses.

The definition of emotion can be broadly divided into three schools: **Perception Theory (James-Lange Theory)**, **Evolutionary Theory**, and **Cognitive Theory** (*Fu, 2016*).

James-Lange Theory tends to study the relationship between emotion and physical changes, and was first defined by James Williams in 1884 (*James, 1884*):

The bodily changes follow directly the PERCEPTION of the exciting fact and that our feeling of the same changes as they occur is the emotion.

James-Lange's Theory emphasizes emotion was followed by physiological action/response, this include internal sensations such as body temperature change and heart pounding, and external behaviors such as crying and running away. The contribution of perception theory is to bridge the gap between emotion and behavior, whether internal physical change or external behavior. Even though this theory seems simplistic at the moment (*Fu, 2016*), because it overemphasized the physical change and ignored the functional influence of emotion on cognitive processes, nevertheless, the theory has inspired future studies of emotion.

Evolutionary Theory emphasizes that the emotion is the result of adaptation in the process of evolution. Human ancestors connected various cognitive functions with each other and formed patterns through adaptation to the environment (*Izard, 1991; Fu, 2016*). The essence of evolutionary theory lies in the triggering of emotions by events themselves, and the corresponding relationship between events and emotions is acquired through learning and genetic coding. Events correspond to the activation of the autonomic nervous system, leading to corresponding physiological responses, as well as a subjective experience and external behavioral manifestations, including expressions and behaviors. In other words, emotion is a coordination mechanism whose evolutionary function is to coordinate various processes in the brain and body in the service of adaptive problem solving, and this series of relationships is built on the basis of natural selection in evolution (*Nesse, 1990; Al-Shawaf et al., 2015; Hammond, 2006*).

Cognitive Theory holds that emotion is triggered by an individual's appraisal of an event (*Lazarus, 1991*). In other words, how an individual interprets the impact of an event on him or her and the extent of the impact determine which emotion is triggered (*Halstead, 1961*). What Cognitive theory and the Evolution theory have in common is that how an event triggers a specific emotion depends on agent's personal emotion experience. For example, without any background knowledge, a man saw the wolf without getting attacked, there was another man who saw the wolf, but got attacked, in the next time when those two people see the wolf, different emotions will be triggered; the man who got attacked may feel fear because he is afraid of getting attacked again, however, the man who didn't get attacked may feel nothing about the wolves approaching. The difference between cognitive theory and evolutionary theory is that the appraisal model of cognitive theory corresponds to a certain kind of event, rather than the one-to-one correspondence between a single event and emotion. Such kind of event will trigger similar emotions, and how to classify those events is

determined by the individual's own experience. Cognitive theory can better explain the difference between different emotions, because it is hard to find the reason why certain emotion is triggered with mapping between event and emotion, but cognitive appraisal can help to show us how the event that triggers emotion impact on the individual.

2.3 Which Theory is Suitable for AI emotion?

In fact, each emotion theory has different emphases due to different research directions, and such differences will also appear in the research of artificial intelligence. Therefore, different artificial intelligence systems may choose different emotional theories due to their different needs, goals and basic assumptions. In the following discussion, I will discuss my choice of emotion theory based on the needs of NARS. It is important to note up front that the following discussion does not represent a good or bad emotion theory; the choice itself is based on the needs of NARS.

James-Lange theory

The focus of **James-Lange theory** is not very suitable for the study of emotion in NARS, because there is only a little discussion about the function of emotion in the theory. Even though one of the functions of emotion is causing external behaviors, but there is not a one-to-one mapping between emotion and behavior, but more complex cognitive processes between emotion and behavior. The structure of the human body and the machine are completely different. If the internal changes of the body are taken as the focus of the emotional research, it is bound to replicate or simulate the internal structure of the human body in the research of the machine. As mentioned above, it is unfair to evaluate psychology's definition of emotion from the perspective of artificial intelligence, and it is also unfair to implement artificial intelligence by completely simulating the structure of the human body.

In fact, the above discussion goes back to the original philosophical question of artificial intelligence, is artificial intelligence infinitely close to human intelligence as the goal, or is it a unique type of intelligence inspired by human intelligence? My answer is more in favor of the latter. Even if AI is implemented in machines, does AI view the world in the same way as humans? If not, copying human intelligence is equivalent to forcing the artificial intelligence to think in the way of human, if, as the essence of artificial intelligence research meaning becomes to create a machine with human thinking, but the result can be solved by means of fertility; why spend time and energy to study and create an intelligent system?

The study of artificial intelligence is a kind of independent intelligence inspired by human intelligence. Similarly, the study of artificial intelligent emotion only has a mechanism similar to human emotion at an abstract level, but this mechanism does not affect how machines will view the world in the future.

No matter what the focus of the research is, all the psychological research on emotion is based on the mature emotional system of human beings. To be more clear, the emotion studied by psychology has already existed, but the research on artificial intelligence emotion is a process growing from nothing.

Evolutionary Theory

Evolutionary theory and cognitive theory have a lot in common, for example, they both establish the relationship between events and emotions through learning, but the differences determine how the relationship between events and emotions is created and expressed. Evolutionary theory emphasis on emotion is the result of human adaptability, in the long-term adaptation, and process of learning, the brain builds events - emotion - coping (physiological reaction, internal change and external behavior), the relationship will be stored in the genetic code to help the agent to handle the same or similar events in the future, or passed on to the next generation.

Although the theory of evolution has been embraced by many emotion researchers (*Barrett, 2017; Ekman, 2009*), it does not, as is often the case with evolution theory itself, offer a clear evolutionary explanation for emotion (*Nesse, 1990*). Instead of giving a more basic, abstract definition of the relationship between the event and the emotion, evolutionary emotion theory leaves evolution and natural selection to account for everything that happens, with specific circumstances as the conditions for the emotion to be triggered.

Even if evolutionary emotion theory is plausible, too specific emotional trigger conditions and irreducible physiological changes in the human body make it difficult to test the theory in an AI system. Because Evolutionary Emotion Theory requires long-term evolution in a “stable environment,” NARS, at least, does not have the cognitive capacity to conduct the kind of extensive learning needed to test the usefulness of Evolutionary Emotion Theory.

Even we ignore the physiological change part of the evolutionary emotion theory and only focus on the event-emotion-behavior clue, and use “Schema” to express the relationship between event-emotion-behavior, it is worth discussing whether such an enumeration approach is consistent with the basic assumptions of NARS design. That is, intelligence is an adaptation based on “relative insufficient knowledge and resources.” And this expression is not “explicable” enough. For example, when people come across a wolf in the jungle, why do some people get scared? However, people with training experience, on the other hand, might not feel fear. What’s the meaning behind this emotional difference?

Cognitive Theory

Cognitive theory emphasizes that emotional triggers have an inseparable relationship with individual cognitive appraisal. Maranon first proposed two components of emotion, namely physiological arousal and psychological or cognitive components.

(*Maranon*, 1985). Arnold judged the positiveness and negativeness of emotions according to the advantages and harmfulness of the event to the agent, so as to take approaches or evasive actions (*Arnold*, 1950).

Unlike other cognitive emotion theorists, Lazarus makes little reference to bodily hormones and physiological arousal in his framework of cognitive theory, focusing instead on the significance of the event itself to the individual. Lazarus believes that emotions are determined and completed by cognitive appraisal, and the triggering of emotions is completely determined by an individual's cognitive appraisal of the event, which explains why the same event will trigger different emotions in different people. The reason is that each individual interprets the same event in different ways.

Appraisal, cognitive regulation, and coping are the three stages of the emotional process (*Lazarus*, 1991; *Stein et al.*, 1990). Emotions are not directly related to things themselves, but the root cause of emotions lies in how individuals understand changes in the environment, in other words, emotion is a certain way of appending the world (*Lazarus*, 1991). The reason why emotion and cognition are inseparable (*Stein et al.*, 1990) is that the trigger of emotion is closely related to the individual's goal (*Lazarus*, 1991), and things unrelated to the individual's goal will not make the individual trigger emotion. The appraisal process is to evaluate the relationship between event or object with individual's goals, so as to determine the type of emotion triggered; the intensity of the emotion corresponds to the extent to which the event affects the individual's goals.

Different with evolutionary theory which link specific events to emotion, the interpretation of the cognitive appraisal method is a kind of abstract way, and appraisal standard is no longer a specific event, but by some of the more abstract parts, for example, Lazarus thinks only events associated with personal goals trigger emotion, and events have nothing to do with goals will not trigger any emotion. This kind of abstract appraisal is also similar to recent research on emotions (*Barrett*, 2017).

The emotion appraisal process discusses a general model for the sources of emotion for most people. The model breaks away from the dependence of emotion on the organism, brain structure, and neuromodulatory system, and only focuses on the interests between the change of external environment and the agent’s “itself”. Such a model gets rid of the constraint that the realization of emotion must involve the neuromodulatory system, and only changes the relationship between the external environment and the “self” so that different emotions are triggered in the “body” of the agent.

As neuroscience advances, neurobiologists are increasingly interested in dissecting the structure of the brain to unravel the sources of emotion. Fewer and fewer psychologists and cognitive scientists are actually talking about emotion appraisal models, and most of the discussion of emotion appraisal models can only be found in the last century (*Lazarus, 1991; Roseman, 1996; Scherer et al., 2001*).

Lazarus’s emotion appraisal model focuses on the direct relationship between the external environment and individual goals, and the appraisal parameters are discrete and nonlinear. The appraisal model is generally divided into two steps. The preliminary appraisal mainly focuses on the relationship between the event itself and the individual goal. In other words, the event which triggers emotion should be related to the individual’s goal, no matter how strong it is. Otherwise, no emotion will be triggered if the event not related to any goals.

The first step also involves checking that the event is congruent with the individual’s goals, which is another way of saying that the event occurs in accordance with the individual’s desires. For example, for an event, the individual’s desire the event to happen. We first ignore the concept of time, so whether the event takes place in the past, present or in the future, then the occurrence of the event is consistent with the individual’s desire, and vice versa (*Lazarus, 1991*). The congruent valence determines whether the emotion triggered by the event is positive or negative.

The valence of congruence determines whether the emotion triggered by the event is positive or negative (*Stein et al.*, 1990; *Lazarus*, 1991; *Ortony et al.*, 1988). When the occurrence of the event is consistent with the individual's desire, the emotion triggered must be a positive emotion, no matter what emotion it is, while when the occurrence of the event is inconsistent with the individual's desire, the emotion triggered must be a negative emotion.

In fact, using the word "positive" or "negative" to describe a particular emotion does not mean that the emotion has a "positive" or "negative" effect on humans. Generally the positive and negative discussion is derived from the emotional experience of emotion and external expression. For example, we used to define angry as a bad feeling, because of the quarrel or fight following anger is not the result of what people want to see. But for the individual, although the cause of anger must be that the individual is hindered in the process of achieving a certain goal, it is positive in terms of the way it is dealt with. Fear involves both avoidance of fear stimuli and a positive attitude towards safety (*Vallverdu et al.*, 2009). By this point, some emotions can be triggered, such as happiness, sadness, and fear, while more complex emotions require more individual cognition (*Lazarus*, 1991).

If a person shows concern or fear about an event, then it probably hasn't happened yet. Even if some people show fear of some object, what they are really afraid of is the harm that the object might do to them, and this fear of harm spreads to the object in relation to the harm, and to all objects that have similar properties to that object. As an example, if a person has ever been attacked by a German Shepherd and developed a fear of German Shepherds, then that fear is likely to spread to all dogs even seemingly docile ones like Samoyed. There is also an old Chinese saying that goes, "Once you got bitten by a snake, you will feel fear of "well rope" for ten year." (Once Bitten, Twice Shy). So, what really triggers fear is the individual's anticipation of possible harm.

Lazarus's appraisal model omits the strength of emotions and the relationship between emotion and time. Emotional intensity affected by many aspects, the first is that how much the event affects individual's goals, for example, a student's expectations on achievement of an exam is 70 points, as long as the result is not less than 70 points, the students will be satisfied with the result, but the satisfaction degree will improve with the improvement of performance, vice versa, if the grade is less than 70 points, the lower grades, the higher degree of sadness the student will experience (*Stein et al.*, 1990).

Time is also an important aspect that affects emotional intensity. Time will affects events differently in the future than in the past. For an event which will happen in the future, besides how much the event itself effect the intensity of the emotion, the closer the occurrence time of the event is to the present time, the higher the intensity of emotional arousal; the further the event is to the present time, the lower the intensity of emotional arousal, it will always be a add-on to the intensity of the emotion because as the time goes by, it will get closer and closer to the occurrence time of the event.

For events that have already happened, the intensity of the emotion decreases over time. What decreased, is the priority of the concept, which refers to the possibility to be revealed by the system in the future working cycles. The decay rate of the priority will be determined by the stability of the concept. The difference between the events that have already happened and happens in the future is, the emotion can be triggered by memories, even though the emotions triggered by the memory have a very low durability, but the intensity of the emotion might be similar when it was happening. It means, even though the system can experience a similar intensity of emotion as before, but the emotion decays really fast once the system realized that the event has already happened. What triggers the emotion is the image of the memory, not the reality (*Damasio*, 1999).

Coping is also a very important concept in Lazarus's theory. As a purely cognitive

approach, and its emphasis on coping rather than physical arousal, laid the foundation for Lazarus's theory of emotion in later research on Functionalist perspective of Emotion (*Joseph et al.*, 1994). The purpose of coping is to change the relationship between the individual and the environment (*Stein et al.*, 1990) in order to maintain the situation that causes the individual to experience positive emotions, or change the situation that causes the individual to experience negative emotions. Its purpose is to maintain the internal environment of the individual, as well as the relationship between the individual and the environment in a state that makes the individual feel relatively stable and comfortable (*Damasio*, 2011). Just as humans learn how to deal with transactions, an AGI system should not pre-define any coping behavior to change system's emotional state. All behaviors require the system to learn during the adaptation process.

Lazarus' theory of emotion is consistent with AGI systems in cognitive function, and the idea of reducing the role of physiological arousal in emotion also provides hope for the realization of emotion in artificial intelligence. The absence of physiological arousal is no longer the reason why machines cannot have emotional functions. Combined with the above two reasons, Lazarus' cognitive emotion theory and Campos' functionalist emotion model become the most important theoretical foundation of this paper.

CHAPTER 3

EMOTION IN AI

3.1 Why Do AI Scholars “Shy Away” From Emotions?

Many psychologists, cognitive scientists, and even neuroscientists have shown us the important role that emotions play in human life from various perspectives. (*Lazarus, 1991; Stein et al., 1990; Bach, 2012; Gadanho, 2003*). Emotion, though difficult to define, but was called by Damasio the most intelligent product of biological values to date (*Damasio, 2011*). But for most of the people, the emotional “disacceptance” of AI comes from the fear of the unknown, that is, the doubt that humans can effectively control the AI system, especially with emotions. This kind of worry mostly comes from the interpretation and presentation of artificial intelligence in science fiction movies, which makes people feel fear of artificial intelligence that have emotions like anger to get out of control.

For AI researchers, the reasons for staying away from AI emotion may be varied.

- **Impossible to create an AI system with emotion**

Some researchers believe that the ultimate goal of AI emotion research, or emotional machine research, is to create artificial intelligence systems with human-like emotions, and the differences between machines and humans make researchers think that such research is absurd and futile (*Ortony et al., 1988*).

- **There is not necessary to have emotion in the system**

Mainstream AI systems have a single task and assume that all time, space, and knowledge are sufficient. This is understandable, since mainstream AI

seeks to optimize the functionality of the product, and therefore to improve the effectiveness of the system at any cost. However, this kind of artificial intelligent system only needs to learn mechanically, and any emotional input is indeed a kind of disturbance to the system. An AI system should simply focus on pursuing the goal it is given or designed for.

- **The AI system is not powerful enough to support emotion functions**

Emotion is an extremely complex cognitive process in the human brain, and no one in psychology, cognitive science, or neuroscience has yet fully understood all aspects of emotion. As for the artificial intelligence system, if its intelligence level is not advanced enough to reach the level of emotion, it is insufficient to instill necessary emotion and expectation into the intelligent system (*La-Grandeur*, 2015). Different from the necessity of emotion, under the condition of limited intelligence or cognition level of intelligent system, the system cannot realize corresponding emotional function even if there is demand for emotional function.

3.2 Why do we need Emotion in AGI

Combining the above viewpoints, we find that the development of artificial intelligence emotion is hindered by multiple factors. The main reason for this problem is that people have high requirements, or expectations, for “emotional machines”. Influenced by the Turing Test, when people discuss artificial intelligence, they will inevitably compare the functions of artificial intelligence system with the same functions in human body. Only intelligent systems with the same abilities as human beings or beyond human abilities will be recognized as “Intelligent”. Therefore, when we talk about emotions, people expect machines to have emotions like humans, and such emotions will be embodied in human-machine interaction, but such requirements

seem to be too harsh for the current research on artificial intelligence.

In addition to the social function of emotion, one of the most important function of emotion is the resource regulation and enhancement of agency autonomy. We often describe a person who is “hard working”, or a “killer”, as “a machine without feelings”. In fact, this sentence reflects from the side that people or machines without emotions are only driven by tasks and goals, and lack autonomy. An autonomous intelligence system should complete tasks in a planned way, and can change the original plan according to the new task in real time. In this dynamic planning module, emotion plays a very important role.

Therefore, in the current work, we pay more attention to the functionality of emotion, that is, how emotion can improve the autonomy of AGI through internal resource allocation, rather than regard emotion as an important regulation of human-computer interaction or machine-computer interaction. Because one of the first things we have to prove is that AGI requires emotion function, in other words, emotion function can improve the reasoning ability and adaptability of AGI.

3.3 Previous Works on Emotion in AGI

The study of emotion in artificial intelligence is difficult to integrate with the study of human emotion in psychology, in part because it is difficult for psychologist to study human emotion without discussing the brain structures that influence emotion (*Damasio*, 2011, 1999), the hypothalamus, the amygdala, which are main structures of emotion, as well as the three major neuromodulatory systems, dopamine, serotonin, and opioid. The examples referenced above, may make people think simulating structure in the brain are associated with emotions and neuromodulatory system is a way to make artificial intelligence system have emotion.

The method of simulation is not necessarily inappropriate. After all, the achievements of deep learning are obvious to all. The neural network imitates a working

mode of human brain in reasoning, even if the machine has no self-awareness. But triggering emotions is inextricably linked to self-awareness, because without awareness, a machine would not know itself exists, let alone know what it is thinking. The appraisal process of emotion is related to self-interest/goal. If the change of external environment is not related to any goal of the individual, no emotion will be triggered (*Lazarus, 1991*). Without the self-awareness, a system that aims only to achieve the function of emotions is hollow, because such an attempt would deprive charm of emotions.

Even artificial intelligent systems that claim to be self-aware (*Li et al., 2018; Franklin et al., 2014; Bach, 2012*) are unable to come up with a logic and method applicable to most artificial intelligence systems due to the differences in the structure of the system, design concepts, and basic assumptions. The fundamental differences in design concept hinders the communication of different system designers. Even if the other party agrees with other design concepts, they will not be able to use the method because it is not compatible with the system they designed.

Many scholars have focused on the external expression of emotion and the corresponding behaviors (*Arbib and Fellous, 2004*), but emotion is also very important for the internal cognitive processes, such as resource-allocation (*Lazarus, 1991; Stein et al., 1990*), decision-making (*Antos and Pfeffer, 2011; Elster, 2009; Gadanho, 2003*), and goal-selection (*Gavrilov, 2008*).

With the continuous development of artificial intelligence technology, artificial intelligence scholars began to explore the possibility of endowing artificial intelligent system with emotion (*Antos and Pfeffer, 2011; Arbib and Fellous, 2004; Bach, 2012; Li et al., 2018*). Although some scholars have denied the possibility of inventing “emotional machine” (*Ortony et al., 1988*), this does not hinder the development of research on emotion in artificial intelligence. Although the difference between human body and machine often discourage artificial intelligence emotion researchers. The

ultimate goal of artificial intelligence emotion research (at least at the present stage) is not to make a machine with the human emotion, but to make the function of human emotion reflected in the AI system.

Any discussion of the possibility of creating an emotional machine cannot be categorically denied because of the structural differences between the human body and the machine. Having emotion modules in an AI system is not exactly an attempt to implement, assign, or simulate human emotion in a machine. I would define the research of emotion in artificial intelligence as an attempt, or as a challenge, because the emotion to the person's enormous influence in our daily life, lets people wonder whether emotion can also effect the machine, or, whether emotional experience can be separated from the structure of the human body and brain, and to be influential in other forms of carriers (animals).

Many different AI researchers have interpreted the role of emotion in AI systems from different perspectives and using different systems. The rest of this chapter will be used to introduce some representative AI emotion research.

Affective Computing

The reason why I put affective computing in the first place is that affective computing is very different from the research direction and basic assumptions of artificial emotion in the general artificial intelligence field. The interpretation of emotion in affective computing and artificial general intelligence is not the difference between strong artificial intelligence and weak artificial intelligence, the main difference comes from the different research direction. In artificial general intelligence system, emotion mainly solves the problem of internal resource regulation, while affective computing solves the problem of human-machine interaction. Therefore, to a certain extent, it is unfair to compare the results of the two research fields. However, the achievements of affective computing in human-computer interaction research can not be ignored,

which will also provide great inspiration for the research of emotion in artificial general intelligence system.

The main application area of affective computing is human-computer interaction. The main purpose of affective computing is to give computers the ability to observe, interpret, and produce emotional characteristics similar to humans (*Tao and Tan, 2005*). Many observations come from the analysis of facial expressions, body movements, gestures, voices, and behaviors. As for the computer, observation is not limited to vision, hearing, or some sensory information used by human beings. It can even scan physiological signals such as human temperature and heart rate through infrared devices, so as to understand each other's emotional state more accurately (*Picard, 1997, 2003*).

Facial expression recognition based on static or dynamic analysis of the expression, the other's feelings make a preliminary judgment, the reason why it is the "preliminary judgment" is because "the facial expression can be deceiving", for example, the face we make when crying. There are several possibilities could make people cry, sadness, "tears of joy," or being moved by something specific, there is only subtle differences of emotional expressions caused by the conditions above.

In addition to facial expression recognition, there are speech processing and body posture, movement recognition. Language processing is mainly based on the content of the speaker, speed, prosodic characteristics and other parameters to analyze and predict the speaker's emotions. Body posture and movement are the collection of emotional information conveyed by body language.

For the above observable information, affective computing adopts a multimodal system, which can analyze all the observable information in a short time uniformly. Externally observable information alone cannot provide a computer with enough knowledge to judge the emotional state of an interactor. According to Lazarus's cognitive affective theory, we know that emotion is triggered by the individual's cog-

nitive appraisal of the event, while expression, language, body posture and movement are all caused by the changes in the autonomic nervous system after emotion is triggered. If you want to interact emotionally with the other party correctly, you need to know the cause of the emotion of the other party. If the person cries because of happiness, the person would share happiness with the other party. The person that is happy should not be comforted by another party because the crying is not for sadness. The affective understanding module of affective computing improves the ability to analyze the emotional state of interactors and optimize their mutual behaviors through the relationship between emotion and cognition.

ACT-R

ACT-R is a typical cognitive structure based on a model-based approach, consisting of a set of modules related to different brain regions in humans (*Park and Myung, 2012*). Each module in the ACT-R corresponds to a similar function in the human brain, including the visual module, the auditory module, and the sound module, which are used for the interaction between the system and the outside world. The Declarative module, the Goal module, and the Imagination module, are used for internal knowledge reasoning and knowledge formation. Some researchers have used ACT-R to conduct related emotional architecture and simulation, so as to solve different problems (*Park and Myung, 2012; Belavkin, 2003; Chown et al., 2006; Belavkin, 2001*).

The realization of emotion in ACT-R is derived from Russell's two-dimensional emotion model (*Russell, 1983,9*). The first dimension is the intensity of emotion, which is related to arousal, while the second dimension is called valence, which is mainly used to simply divide emotion into two dimensions, positive (positive emotion) and negative (negative emotion). Choice in ACT-R is not only influenced by goal (G) and noise (T), in addition to the experience stored as rules, which are the primary

factors of emotional arousal. The ACT-R has different motivations, arousal levels, and coping styles for different events, depending on the goal and different level of noise. Its main functions are used to resolve conflict issues and help the system make choices.

Belavkin also demonstrated the need for emotion in the cognitive system in his doctoral dissertation (*Belavkin, 2003*), demonstrating the effect of emotional stimulation on learning efficiency in mice by replicating Yerkes and Dodson’s “Dancing Mouse” experiment (*Yerkes and Dodson, 1908*) in ACT-R. The definition of conflict in ACT-R is mainly shown as the process of selecting one rule from multiple rules matching multiple goal states, and the addition of emotion makes the agent more inclined to behavior selection in the selection decision-making process. For example, in the case of low noise (T), ACT-R produces positive emotions and high self-confidence, corresponding to behaviors that tend to be risk-averse in order to achieve beneficial effects, whereas in the case of high noise, choices tend to be risk-taking and carry negative emotions. When the goal value is low, the arousal is not high because of the stimulus degree, the awakening degree is low, and then the performance is lack of motivation, the effort to complete the goal is small, and for the high goal value, the agent feels positive, the awakening degree is high, and is willing to spend more energy to complete the goal. As a result, Belavkin also described the effect of the ratio of G to T on the agent by combining the effects of goal and noise, and found that either too high or too low a stimulus would hinder learning and reasoning, while moderate stimulus would promote certain cognitive functions.

CLARION

CLARION is a cognitive structure composed of multiple subsystems, including: action-centered (ACS), non-action-centered (NACS), motivational-centered (MCS), and metacognitive (MS) subsystems(*Sun et al., 2015*). Emotion in CLARION is

regarded as the synthesis of multiple mechanisms and processes, including experience, behavior, cognition, psychological and physiological characteristics. It is for this reason that emotion in CLARION is the result of coordination of multiple subsystems. CLARION's synthetic cognitive framework makes the realization of emotion in its system possible. The emotional process is also divided into explicit and invisible processes, which interact with each other and play the role of cooperative reasoning in decision-making and action.

CLARION's principle of reasoning is that action comes first, and reasoning serves action. This makes CLARION adjust CLARION's attention through emotion all the time, and further influence CLARION's choice of action. CLARION's affective model includes: reactive affective, deliberative assessment, and coping/action composition. Reactive Emotion and Prudential Assessment (Reactive Emotion and Prudential Assessment) identifies the resulting emotion by conducting a detailed assessment of the current event in relation to the benefits of the system itself. The results can be used in the third step, response/action selection.

Emotion is considered to be directly related to action in CLARION, and it is claimed that emotion is expressed through action to a large extent, and emotion has a great influence on perception and attention. While inside the system, emotion is generated through the appraisal process jointly executed by ACS, NACS, MCS and MS. Each subsystem is evaluated at different levels and in different ways. Prudential appraisal can be carried out in NACS, and a goal is recommended, and the goal determines the coping/behavior. Goals will be selected by the ACS and simulated, compared, and summarized.

LIDA

The LIDA model is a conceptual and partially implemented computing model, the design of the system is based on Barrs' global workspace theory (*Baars, 2017*), which

emphasizes the role of consciousness (in particular attention) in cognition. In simple terms, the LIDA can do real-time interaction with the environment, to understand the environment data, and various modules within the system will be further filtering and use these to understand, and compete in the global workspace for attention, and the content of the victory will be broadcast in the global working space, while the conscious broadcast will further recruit some unconscious process, eventually, it will choose an appropriate response based on the current situation (*Franklin et al.*, 2014).

Emotion is emphasized as a feeling in LIDA, and feeling is realized in internal Perceptual Associative Memory (PAM) nodes. According to different valence states of feeling (positive or negative), feeling nodes are divided into driving feeling nodes and interpretive feeling nodes. The former describes the current internal state of the agent, while the latter is used to evaluate internal or external stimuli (*McCall et al.*, 2020).

The emotional valence state determines LIDA's attitude towards an object's likes or dislikes, while activation determines the degree of emotion. Activation itself also represents the salience of the situation. Therefore, the activation state corresponding to emotion contributes to the competitive strength of relevant nodes in the global workspace, and makes the nodes with high activation state more likely to win the competition for attention.

LIDA also uses Lazarus' cognitive emotion appraisal theory as the system itself to evaluate the relationship between events and the system itself from the perspective of emotion. Assessment, as the initial pre-conscious process, is triggered automatically by perception, but it can also be triggered by memory and imagination.

MicroPsi

MicroPsi is an agent architecture based on Dietrich Drner's Psi theory, which attempts to describe the emotional, motivational, and cognitive interactions of situ-

ational agents in artificial intelligence systems (*Bach, 2003*).

MicroPsi is a structured implementation of the Psi theory, similar to LIDA, which is a structured implementation of the global workspace theory. MicroPsi focuses on the agent’s own motivations, emotions, and the acquisition of objects in the environment in the process of interacting with the environment, as well as the interaction with objects. The MicroPsi is connected to the environment through a set of body parameters and external sensors, and all the semantics of acquired representations are the result of interactions with the environment (*Bach et al., 2019*).

Emotion in MicroPsi is closely related to needs, satisfying a need produces a happy signal, and conversely, a pain signal. For a goal, if it is pursued by the agent, a positive reward will be generated to urge the agent to pursue the goal; otherwise, a negative reward will be generated. In order to avoid the event.

Emotions are not explicitly realized in MicroPsi, but appear as perceptual classifications, describing 32 emotions according to different needs, different environments the agent is in, and different relationships between the agent and the environment.

Sigma

Sigma is a cognitive architecture and system whose development is driven by a combination of four desiderata: grand unification, generic cognition, functional elegance, and sufficient efficiency, it is built around an eponymous cognitive architecture: Sigma (*Rosenbloom et al., 2016*). The previous Sigma system did not have the emotional mechanism, but in order to strengthen the ability to deal with complex real-world problems, expand the ability of the general mechanism, and improve the efficiency of the system to solve problems, Sigma tried to integrate the emotional mechanism into the system (*Rosenbloom et al., 2015*).

Even though the earlier Sigma did not have emotional capabilities, the overall architecture of Sigma fully supports the capabilities necessary for emotional processing.

The designers of Sigma believe that emotional modules should not be independent of the existing functions of the system, but to implement the architecture of emotional modules through the original functions as much as possible. The emotion module mainly contains two functions: (1) the basic appraisal of emotion trigger; (2) the regulation of thought and behavior after emotion trigger.

Sigma's emotion model takes into account the impact of anticipated events on emotions and the desirability of the event, i.e. whether an event promotes survival and hinders the desired goal, and at the same time, the primary function of emotions is defined as the efficient allocation of attention to limited resources.

CHAPTER 4

NON-AXIOMATIC REASONING SYSTEM

4.1 NARS Overview

NARS (Non-Axiomatic Reasoning System) is an AGI designed in the framework of a reasoning system. The project has been described in many publications, including two books (*Wang, 2006*)(*Wang, 2013*), Therefore, this chapter only introduces the content that is directly related to the doctoral dissertation.

Theoretical and strategic assumptions

Mainstream AI defines “intelligence” as the ability to solve problems that only the human brain can solve. Obviously, such a definition limits AI to the framework of “human intelligence”, but this definition doesn’t seem fair to intelligent machines. For example, airplanes are designed with birds and other animals in mind, ”it is clear that airplanes have designs and capabilities different from birds. Although the scope of problem-solving of current intelligent systems cannot be compared with that of human beings, it is undeniable that computer systems are indeed better than human beings in terms of computational speed and accuracy, as well as memory. But at the same time, it does not mean that the ability of intelligent system will surpass human beings. There is an insurmountable gap between human beings and machines, and one side is always more advanced than the other in some regard. If the advantages of intelligent machines are properly utilized, intelligent machines will eventually become a good helper for human beings.

In NARS, intelligence is defined as “*the ability for a system to adapt to its environment and to work with insufficient knowledge and resources.*” This definition requires the system to be able to perceive unanticipated information from any environment, working with finite resources, interact with the environment according to experience and finally adapt to the environment. At any given time, an action performed by NARS while interacting with the environment or the answer the system gives to a problem is not necessarily the optimal solution to the problem, but the best answer that can be given based on the experience of NARS. Therefore, as NARS continues to learn and adapt to the environment, its ability to deal with problems will gradually improve.

According to the NARS definition of intelligence, a NARS system has three basic properties:

- **Finite:** The system has a constant information-processing capacity in terms of processor speed, storage space, etc.
- **Real time:** The system is capable of handling problems that arise at any time, and the utility of its solutions may decline over time.
- **Open:** The system can accept input data and questions for any content, as long as they are presented in a format recognized by the system.

These properties determine that when NARS is working with limited resources and knowledge, it is not possible for NARS to consider all solutions to a problem, but rather to consider important, more relevant possibilities, and then make inferences based on past experience to come up with a “relatively rational” solution.

Knowledge Representation

As a reasoning system, NARS uses a formal language called “Narsese” for knowledge representation, which is defined by a formal grammar given in *Wang (2013)*.

This chapter only introduces the knowledge representation related to this dissertation, so that readers can understand the following introduction.

The logic used in NARS belongs to a tradition of logic called “term logic”, where the smallest component of the representation language is a “term”, and the simplest statement has a “subject-copula-predicate” format, where the subject and the predicate are both terms.

The basic statement in NARS is the inheritance statement, which takes the form of “ $S \rightarrow P$ ”, where S is the subject term, P is the predicate term, and the “ \rightarrow ” is the *inheritance* copula, which is defined as a reflexive and transitive relation from one term to another term. The intuitive meaning of “ $S \rightarrow P$ ” is “S is a special case of P” and “P is a general case of S”. For example: statement “ $water \rightarrow liquid$ ” intuitively means “water is a type of liquid”

In general, the “inheritance” copula is similar with the subset relation in set theory (Wang, 2006). “ $robin \rightarrow bird$ ” intuitively means “Robin is a type of bird”, not robin is a particular instance of a bird. But, if there is a robin named Tweety, then Tweety is an instance of robin as well as a bird. “Tweety is a robin” is written as “ $\{Tweety\} \rightarrow robin$ ”. In addition to instances, attributes of an object can be represented as “[*Term*]”. For example, “ $\{Tweety\} \rightarrow [yellow]$ ” intuitively means “Tweety is yellow”.

Atomic terms can be combined in Narsese to construct *compound terms* of various types. A compound term ($con, C_1, C_2, \dots, C_n$) is formed by a term connector, *con*, and one or more component terms (C_1, C_2, \dots, C_n). The term connector is a logical constant with pre-defined meaning in the system.

The way in which a sequence of events can be represent in a format of compound term, for example, a sequence of events, A, B, and C can be represented as a compound term ($\&/, A, B, C$). It means event A, B, and C happen one after another. The sequence does not have the same meaning as the Narsese “implication”, which means

“if-then””; for example, $E_1 \Rightarrow E_2$ means if E_1 happens, then E_2 will happen.

The above statement can be regarded as being used by NARS to describe an object or event in the third person view. In addition to external description, Narsese can also describe the internal activity of NARS through a special Narsese statement called an “operation ”, which is an event that can be directly realized and executed by the system.

In general, an operation statement in NARS is expressed as:

$$(\wedge_{\text{op}}(a), \{\text{SELF}\})$$

This means that NARS will execute operation a . And $SELF$ is a special concept in NARS, referring to NARS itself. When $SELF$ appears, NARS is no longer objectively described from the third-person perspective through Narsese, but describes NARS’ own subjective feelings, subjective experiences and triggering subjective operations in first person form through their relations with the $SELF$ concept.

In general, there are three types of sentences in Narsese:

- A **judgment** is a statement with a truth value, and represents a piece of knowledge that system knows or can learn. For example,

$$\langle \langle \&/, \langle \{\text{enemy}\} \rightarrow [\text{left}], (\wedge_{\text{left}}, \{\text{SELF}\}) \Rightarrow \langle \{\text{SELF}\} \rightarrow [\text{good}] \rangle \rangle \rangle.$$

means when the enemy appears on the left and the NARS moves to the left, the NARS will be good. This sentence with a truth-value makes the system able to absorb this conceptual relation, together with its implications, into the system’s beliefs. More details about the truth value can be found in (*Wang, 2006*).

- A **goal** is a statement to be realized by executing some operations. For example, if NARS has a goal “ $\langle \{\text{SELF}\} \rightarrow [\text{good}] \rangle!$ ”, means that the system should keep

itself in “good” condition, and if considering the example above, one way to keep itself in “good” condition is to move to the left when the enemy appears on the left. Each goal is accompanied by a desired value, which represents the degree or state of the events that the system wants achieve. More details about the desire value can be found in (*Wang, 2006*), too.

- A **question** is a statement without a truth-value or desire-value, and represents a query to be answered according to the system’s beliefs or goals. For example, if the system has a belief “ $\langle \{SELF\} \rightarrow [good] \rangle$.” (with a truth-value), it can be used to answer question “ $\langle \{SELF\} \rightarrow [good] \rangle?$ ” by reporting the truth-value, as well as to answer the question “ $\langle \{SELF\} \rightarrow ?x \rangle?$ ” by reporting the truth-value together with the term $[good]$, as it is the property of $SELF$. Similarly, the same belief can also be used to answer question “ $\langle ?y \rightarrow [good] \rangle?$ ” by reporting the truth-value together with the term $\{SELF\}$, as it is the instance that has property $[good]$.

4.2 Inference control

According to the inference rules of NARS, NARS can perform the following three inference tasks:

- To absorb new experience through interaction with the environment. If it is new experience, it will be added to the beliefs of the system. If there is already the same belief in the system but the new and old beliefs have different evidence base, the new and old beliefs will be combined through “revision”. In addition to revision, it will also make use of the original knowledge to derive some of their implications spontaneously.;
- To answer input questions and derived questions according to the system’s beliefs.

- To achieve input goals and derived goals by executing the related operations under corresponding context according to the system’s beliefs;

In NARS’ memory, beliefs and tasks are organized into *concepts*, and all the concepts are put into a data structure named “bag”. A “bag” is a special data structure designed for resource allocation in NARS. All the concepts in the bag are grouped into different levels according to their priority. High-priority concepts are allocated more resources for processing, and similarly, low-priority concepts are allocated fewer resources. Each working cycle of NARS selects a concept from the bag to be processed, and high-priority concepts have a higher chance of being selected, while low-priority concepts still have chance of being selected, but are less likely to be selected than high-priority concepts. Therefore, the probability for a concept to be selected is positively correlated to its priority value.

Attention is mainly determined by priority and durability, which are part of a “budget value” (*Hammer et al.*, 2016). In general, priority corresponds to how important a concept or task is to NARS, so the more important the concept or task is, the greater the chance that it will be processed first by NARS. Durability corresponds to decay rate of the priority; the priorities of concepts and tasks will decay over time. The higher the degree of durability, the priority will decay slower, the concept or the task can be focused on by the system for a longer time, on the contrary, the lower the degree of durability, the priority will decay faster, the concept or the task can be ignored faster by the system.

4.3 New Architecture

My work on emotion contributes to the implementation of a new architecture of NARS, which was introduced into the system initially in OpenNARS version 3.1.0 (*Wang et al.*, 2020). This section will give a brief description of the framework of the

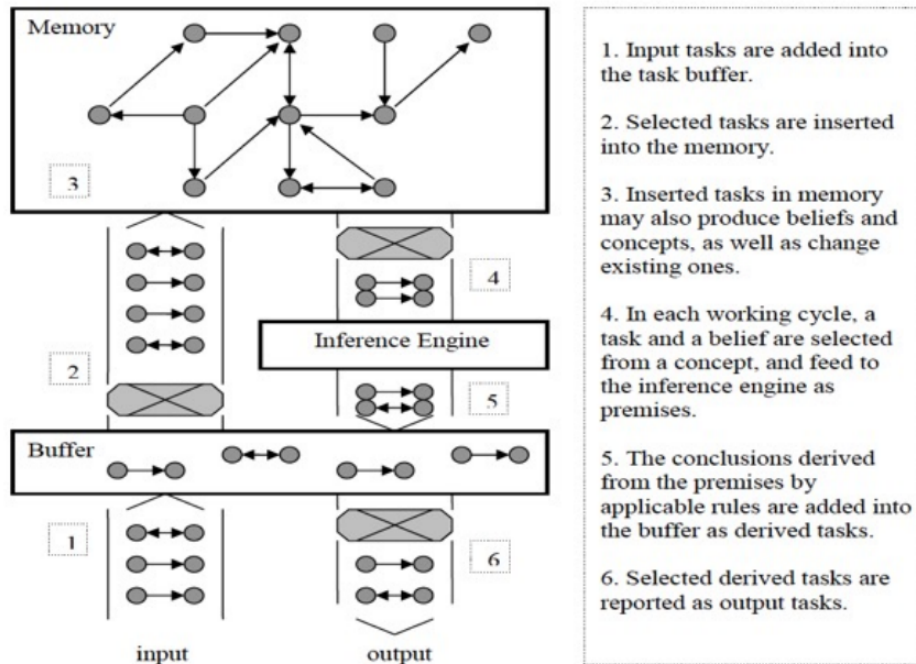


Figure 4.1: The old Architecture of Control Mechanism in NARS

new architecture, while the more specific implementation and functions within the architecture will be introduced in the following sections.

Compared with the old control mechanism, the new control mechanism of NARS is mainly reflected in the change of buffer. Figure 4.1 shows the old control architecture, which was largely used in previous releases. In the earliest versions, buffers were simply containers for carrying input, since the previous tests did not have large or frequent inputs, and many of the test sets were simple one-step reasoning. When different numbers of inputs occur at the same time, all of them are stored in the buffer, waiting for the next working cycle to start before sending all the inputs into memory in turn. This approach is obviously in conflict with the basic assumption of NARS, “adaptation with insufficiency of knowledge and resources (AIKR)” to some extent.

Problems in Previous Versions of OpenNARS

Under the assumption of AIKR, it would be unreasonable for NARS to process all incoming information, unless NARS always has enough time and space to process all the information receives; but it is clear that real life does not confer such conditions on any person or machine, any time and space are not infinite, we cannot discuss “infinite” possibilities in a “limited” framework. If NARS does not have such a capability, extracting all the information from the buffer in each working cycle will cause great pressure on NARS reasoning, especially when there is a large amount of information flooding in continuously, NARS does not have enough time and space to process all the information in a short time and continue to receive new information.

In later versions(*Hammer et al.*, 2016), the buffer functionality was updated. Instead of taking all the tasks from the buffer at a time, the main memory selects only one task from the buffer at a time. If the selected task is an event, temporal induction will happen between the selected event with all other events in the buffer right before or after it.

The buffer design is still flawed in this version. Even though the buffer capacity is limited, the time that an event exists in the buffer tends to be infinite. Such a setting can create redundancy in the buffer when a large number of events rush in, that is, the event itself will only be removed by the buffer when the buffer capacity is reached and the priority of the event itself is the minimum in the buffer, otherwise the event will remain in the buffer until it is selected. Such a design would obviously conflict with the basic assumptions of AIKR to some extent, and would prevent later events from being perceived by the system if the event remained in the buffer.

Another problem is the single buffer setup. The single buffer accepts both external information and internal reasoning results, which causes information redundancy and fails to provide sufficient support for the emotion module. In short, a single buffer cannot enable the system to have an emotion module capable of handling both

external and internal stimuli.

The last problem is the inability to deal with anticipations in a timely manner. As a very important function of the cognitive system, anticipation can make a prediction of what will happen next after an event is perceived by the system according to its own experience. For example, the system may know that B will happen after A, so when A happens, the system will expect the occurrence of B. But the previous version could not predict the occurrence of B at the same time as A occurs and is perceived by the system, but only can make this prediction after A has entered the main memory. In the face of two events with short time interval, processing cannot make a timely prediction. Sometimes, the prediction will be generated after B has happened, and such result is bound to affect emotions related to the anticipation. Fear, for example, will not be triggered if the system is unable to anticipate consequences in the face of danger, because fear is about something that has not yet happened, rather than something that has already happened. Delayed anticipation triggering can cause the system to be harmed by the inability to anticipate a bad outcome, even in the face of danger.

While not all of the problems have been covered in the previous discussion, the issues mentioned are the ones that must be addressed to build emotional modules. The rest of this chapter discusses the new control architecture for NARS and how the new architecture addresses these control issues.

Overview of New Architecture

This section briefly introduces the framework new architecture of NARS. The new architecture will update the number and functionality of buffers on the basis of the old architecture, update the location of the anticipation modules while addressing the internal problems of the previous version, and add emotional functionality on the basis of the new architecture. In the new architecture, emotion is not programmed

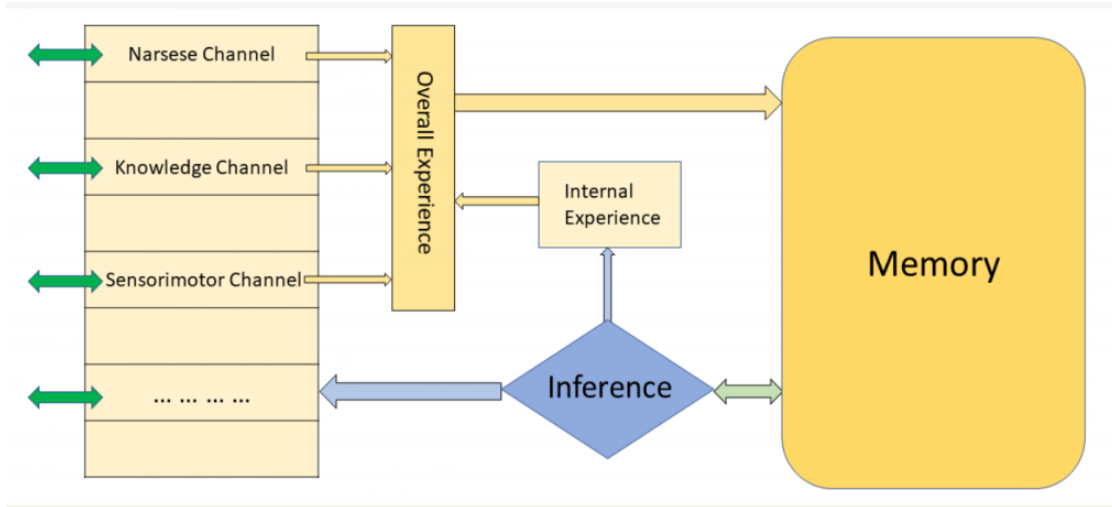


Figure 4.2: The New Architecture of Control Mechanism in NARS

as a plug-in or an additional function of the system, but is integrated into the basic cognitive processing and becomes an essential part of the cognitive process.

The new architecture will adopt a “double buffer” construction. As shown in Figure 4.2, the buffer is divided into two parts, the Overall Experience buffer and the Internal experience buffer, while keeping the main memory and the inference module unchanged.

Overall experience buffer is used to store and process both external and internal stimuli. The external stimuli are perceived by the system. Although current systems are not perfectly equipped with sensors connected to the outside world, this does not affect the creation of Overall experience buffers. The internal stimulus comes from the internal experiential buffer. The internal buffer only stores and processes internal stimuli, and the inference results of each working cycle will be first sent to the Internal Experience buffer for processing instead of being directly sent to the Overall Experience buffer.

The overview of NARS shows that the capabilities provided by NARS have the potential to build evaluation models for some emotions.

The goal processing function of NARS itself provides the possibility to realize the

emotional process in NARS. First, when an event occurs, whether the event itself is relevant to any of the NARS goals determines whether an emotion will be triggered. If the event related to any of the goals, the emotion module of NARS will be triggered. By this point, some emotion will be triggered but we don't know what type of emotion will be triggered.

The congruence between the event and the goal determines the valence of the emotion. Positive emotions are triggered when the event is congruent with the goal, while negative emotions are triggered when it is incongruent with the goal. If the event is realized to be congruent with the desire of NARS (desire value is greater than 0.5), the positive emotion module will be activated and further specification will be required, other wise, if the event is realized to be incongruent with the desire(desire value less than 0.5), the negative emotion module will be activated and further specification will be required to decide which emotion NARS is experiencing at the current moment.

Cognitive regulation is a bridge connecting appraisal and coping. Cognitive regulation adjusts the priority, durability and quality related to the goal concept within NARS according to the result of appraisal (what kind of emotion is triggered, and the intensity of the emotion) , so as to improve/reduce the possibility of the related concept being processed by NARS, as well as the decay rate, thus affecting the strategies and plans of NARS for dealing with related events.

More detailed introduction and discussion about the principles of the new architecture, as well as the internal architecture and functions of Buffer will be found in chapter 7.

CHAPTER 5

REQUIREMENTS FOR AN AGI EMOTION MECHANISM

This chapter describes the requirements for an AGI emotion mechanism. This discussion will be developed from several different aspects, based on the existing NARS emotion model, to describe the conditions necessary to implement emotion model in AGI system. The discussion on this issue will be carried out from the following aspects: a). **Design Requirements:** this part mainly discusses the design requirement which is the theoretical basis of AGI emotion model design. The framework of emotion model determines the how it can be built based on the current system. The design of emotional modules must be based on the design framework of the current system to make it possible to be implemented, rather than a idealized theoretical framework that cannot be implemented. b). **Functional Requirements:** Functional requirements mainly discuss what changes emotion will have on the function of AGI system. AGI researchers cannot add emotion to the cognitive framework simply because emotion is an essential part of human cognitive process. They cannot add emotion just for the sake of adding emotion. The purpose of adding emotion modules is to improve the performance of the cognitive system, whether it is efficiency or system autonomy. c). **Architectural Requirements:** The architectural requirements mainly discuss the architectural requirements of the emotional modules if they are added to the AGI system, no matter whether the emotional modules are integrated into the functions of the system or an independent module separate from other functions of the system.

5.1 Design Requirements

Emotion, as an indispensable part of human cognition, is a hot research field in psychology, cognitive science and even neurobiology. Emotion has also been favored in the field of artificial general intelligence. Through the discussion in the second chapter, we find that there are great differences between the mainstream AI field and the AGI field in the attitude towards emotion, the basic assumptions for the implementation of an emotion machine, and the methods for the implementation of an emotion machine. However, the difference on fundamental assumptions does not affect our discussion on this issue. Mainstream AI study for emotional artificial intelligence began in the discussion and implementation around human-computer interaction, which is feature analysis of people's emotional state, these features include, expression, sound, body temperature, even the experience of the people which triggered the emotion, so as to make the right feedback. In comparison, the artificial general intelligence began in the impact of emotion on system's function which is based on the analysis of the current event or relationship between environment and system, which trigger some emotions and lead the system into different emotional states. The emotional state will adjust internal environment and the system's parameters so it can make some response; this kind of response is either internal ideological changes or changes in the external behavior.

Design requirement 1: Emotion should not be considered as an additional function of the cognitive system, but should be integrated directly into the cognitive framework

The requirement does not mean all cognitive architectures need to have an emotional module, but once a cognitive framework needs to add an emotional module, the

module should not serve as a tool or plug-in which independent from the overall cognitive architecture, not should it be added in order to solve one specific problem within the cognitive architecture. Emotions are supposed to be the result of how the whole cognitive system works (*Sun et al.*, 2015), and the result will act on the whole cognitive system. The addition of emotion is a comprehensive improvement of the ability of the cognitive system, which is an extension of the strict cognitive processing to improve the intelligent behavior needed in the complex real world(*Rosenbloom et al.*, 2015). Therefore, the design of emotion modules is not only about how the current architecture can support the elements necessary for emotional processing, but more importantly, how emotion can be integrated into the existing cognitive architecture.

Most AGI systems follow this design concept in the design of emotion module, and the new architecture of NARS system also follows this design requirement. In the new architecture, the emotion module breaks away from the design concept of NARS 3.0.4 which make emotion module as a plug-in, but instead integrates different emotion assessment, regulation and feedback into the overall architecture. At the beginning of the new architecture design, emotion has been regarded as an essential component.

The trigger of emotion is not only due to change of external environment, but also the own thoughts and memories of NARS. The emotion module in the new architecture is not an independent module from the other modules, instead it's in a different positions which have the corresponding processing mechanism. For example, a relatively simple emotion, fear, in simple terms, the appraisal model for fear is when an event expected by the agent is contrary to one of goals of the agent (*Lazarus*, 1991). Therefore, fear is related to the anticipation, and the appraisal model and regulation function of fear are emerged into the module of anticipation processing

Design Requirement 2: The emotion modules in the AGI system should

be constructed in accordance with the elements provided by the cognitive structure.

If the AGI system is classified according to the theoretical basis, we can simply divide it into two categories: one school tries to accurately simulate human emotion (*Bach, 2012; Franklin et al., 2014*), while the other try to realize a more general concept of emotion that is not limited to human emotion (*Wang, 2006; Belavkin, 2003*). This paper is not intended to evaluate the above two types of cognitive architecture. In fact, no matter which type of cognitive architecture is used, a complete theoretical basis is required for the implementation of emotion modules. The theoretical foundation is not simply to provide theoretical guidance to the emotion module, but more importantly to detect whether the cognitive framework has reached the design level that can realize the corresponding emotion model.

The research on emotion in the field of psychology, cognitive psychology, and even neurobiology research are in the opposite direction of research in the field of artificial intelligence. The research of emotion in psychology is based on the emotion already existing in human cognitive structure, while the research of emotion in artificial intelligence, especially in AGI research field, is inspired by human emotion and creates or explores similar emotion models in AGI system from nothing. Most cognitive architectures may not have emotion models in the early stage.

In the discussion in chapter two, we aimed at several different psychological emotion models, the reason why cognitive emotion model is selected as the theoretical foundation for the design and implementation of emotion module is because cognitive theory has completely described the process from emotion triggering to how emotion plays a role in people's cognitive function, taking emotion as an important role in the cognitive process. Emotion evolution theory focuses more on how the emotional mechanism is established, while James-Lange theory only emphasizes the mapping of

emotion and behavior as well as physiological changes, but ignores the influence of emotion on people at the cognitive level. In other words, when a behavior is accompanied by emotion, James-Lange theory does not attempt to explain how emotions drive behavioral or physiological responses. Therefore, based on the cognitive emotion theory, it is possible to build a complete emotional framework in NARS. In addition to NARS, Micropsi, LIDA, Sigma and other cognitive frameworks have also established emotion modules based on different emotion theories.

The only thing that is certain is that even if the evolutionary emotion theory is reasonable, most AGI systems are not able to obtain emotions through evolution at this stage. As a part of evolutionary theory, evolutionary emotion theory has strict requirements on a stable evolutionary environment, which can not be provided by AGI system today. Moreover, many social emotions can only be generated by the communication between multiple individuals. At least the current realization of NARS cannot build emotion module on the theoretical basis of evolutionary emotion theory.

To sum up, the realization of emotion module in an AGI system should rely on the elements which can be provided by the existing cognitive framework, rather than on the theoretical model that the framework does not support.

Design Requirement 3: Emotion modules should be geared towards information from different sources, rather than specific sources.

The main idea of cognitive emotion theory is that emotion arises from the individual's cognitive appraisal of events. Individuals produce corresponding emotions by evaluating the impact of events on themselves. But in real life, emotions are triggered by different channels, auditory, visual, and tactile. The non-event information such as knowledge may also trigger emotions like being surprised or disappointment just

because the new experience is different from the old knowledge.

Therefore, when the emotion module is considered to be added to the cognitive structure, the information that the module is faced with should be comprehensive, not just for a particular piece of information. Of course, different emotions are sometimes triggered independently of other emotions. For example, emotions that are associated with expectations are not triggered by events that have happened in the past, and emotions that require certainty of fact are not triggered by events that have not yet occurred. But it can't be that emotion should face the need for information from all kinds of channels, because emotion is a special state that runs through the whole cognitive framework.

Design Requirement 4: Emotions cannot only be triggered by external events; reasoning results within the system, memories of past events, thinking and imagination should all have the ability to trigger emotions.

External events for individual emotional trigger is just one of the many possibilities, the sometimes invisible cognitive process of human brain also trigger corresponding emotions, sometimes reasoning, conception, and fantasy are likely to trigger a corresponding emotion, and through these cognitive processes emotion also has a very important role. For example, when a person plans to do something, the anticipation of the outcome may trigger the corresponding emotion, which promotes or prevents the action from being further carried out, and which promotes or prevents may depends on the positiveness or negativeness of the emotion that triggered by the anticipation of the outcome.

The human capacity to empathize is also to put ourselves in the other person's situation to feel the other person's emotional state, which is the result of a construct rather than what is actually happening to us. This kind of conception can place oneself

in the other party's situation and experience the other party's feelings from the other party's point of view, so as to trigger the corresponding emotion in oneself to achieve the purpose of empathy, and strengthen the effect of communication. Although as mentioned above, human-computer interaction is not the main development goal of AGI system at the present stage, it does not mean that human-computer interaction will not be studied in AGI in the future. However, effective communication is based on mutual understanding between the two parties. If the AGI system can trigger emotions only because of what happens to itself, then it is unlikely to achieve good effects in human-computer interaction in the future.

5.2 Functional Requirements

Unlike affective computing, only a small number of AGI systems implement or focus on human-computer interaction (*Barone et al.*, 2008). The main purpose of the emotion model in most AGI systems is to change the original reasoning mode, improve the efficiency of reasoning, enhance the versatility, and improve the autonomy of the system by using the function of emotion. Instead of analyzing human emotions and giving correct feedback, the AGI system is inspired by human emotions, allowing agents to solve problems in the complex real world through emotional functions just like human beings. The emotional function here is not necessarily to express the individual's emotional state through facial expressions and actions, so as to make the communication more clear and direct, but to change the internal knowledge structure through emotions, so that the system can reason and work selectively.

Functional Requirements 1: The emotional mechanism in the AGI system requires the ability to detect goal-related events, whether actual or expected by the anticipation. This mechanism should be generic, not spe-

cific type of events which are predetermined.

Lazarus’s cognitive emotion theory first emphasized that the events which can trigger emotions must be relevant with the goal of the agent (*Lazarus, 1991*), whether the event itself is associated with the goal, or due to the occurrence of an event, the individual expectations of another incident related to the goal: such events will trigger emotions, and events not related to goals will not trigger the emotions. The functional mechanism of emotion and goal association is also used in many AGI systems with emotion mechanism (*Franklin et al., 2014; Rosenbloom et al., 2015; Belavkin, 2001; Bach, 2012*). The association with goal can be used to filter out the events unrelated to the emotion, so that the events unrelated to the goal are not involved in emotional processing.

From the above discussion, if the emotion is triggered by an event associated with a goal, then for an AGI system, the AGI system must have a **Goal** processing mechanism. Without the ability to detect or process the goal, the system cannot even make a preliminary judgment whether an event will trigger emotions. In this case, simply rely on “implication” which mapping the event and the emotion will make AGI system losing generality to a certain extent, exhausting all possibilities requires a lot of resources, energy, and time, it contradict to the basic assumption that the AGI system was designed under the assumption of “insufficient knowledge and resources (*Wang, 2013*).”

Functional Requirements 2: The emotion mechanism of AGI system should have the function of internal resource regulation.

As mentioned in the previous discussion, AGI systems deal with emotions differently than mainstream AI. Affective computing is mainly used for detecting another

agent's (mainly human) emotional state and making the corresponding feedback. But what is realized in the AGI systems is human-like intelligence, not human-level intelligence. The intelligence in AGI is inspired by human intelligence, and tries to achieve similar intelligent process in the machine; in this process, the advantages of machines can be reflected, and may even surpass human intelligence to a certain extent. The idea is not far-fetched, as Deep Blue and Alpha-Go have already shown us it is possible.

The biggest difference between humans and machines cannot simply be described as humans having emotions and machines not having emotions. The biggest difference between human and machine lies in whether there is autonomy: human can think on their own, can take action according to their own consciousness, can choose the goal from different goals to complete, but the computer is always run in accordance with the pre-set goals and operations.

The goal of AGI systems is the ability of the system to deal with complex problems in the real world, which are often real-time and sometimes unpredictable. When the machine faces the real world, it tends to receive a large amount of information at the same time, rather than a specific information. Faced with such a large amount of information, the AGI system should have the ability to actively filter the information that is worthy of being received by the system. Similarly, it should have the ability to autonomously conduct internal resource control to decide which tasks to prioritize or which tasks to forgo.

One of the important functions of emotion is to arouse. It can arouse certain goals when necessary, so as to prompt the agent to focus their attention on these goals. The valence of emotion decides whether to promote the occurrence of certain events or try to hinder the occurrence of certain things. For example, when emotion "hope" is triggered, the agent will focus on anything that can make the goal happens, while when emotions such as "fear" is triggered, the agent will focus on how to avoid the

possible occurrence of bad things.

Emotion does not directly produce behavior, but generates motivation to accomplish a specific goal by diverting attention. Therefore, there is no one-to-one mapping between emotion and behavior, and previous studies have also found that the establishment of the relationship between emotion and behavior by “implication” does not play an significant role in reasoning (*Li et al.*, 2018); even removing the information containing emotion will not affect reasoning.

5.3 Architecture Requirements

Lazarus’s theory of emotions describes an emotional process as being divided into three stages: appraisal, arousal, and coping. Appraisal is to evaluate the relationship between the events that are happening or about to happen (including events) and the agent according to the environment the agent is in. The appraisal is further divided into three levels. First, the correlation between the event and the goal is evaluated. Those related to the goal are further processed, while those unrelated to the goal are not further processed. Arousal refers to internal resources and knowledge regulation and change based on the appraisal results. The intensity of regulation is determined by the degree of arousal. The higher the degree of arousal, the greater the intensity of regulation, and vice versa. Coping is the response to arousal. Coping can be either unconscious or conscious.

According to the above discussion, it is not difficult to conclude that a complete emotional architecture should also be able to achieve the above three processes.

Architecture Requirement 1: The emotional architecture in AGI system should have the function of appraisal, and it needs to be generic.

The importance of cognitive appraisal function for emotion modules is not only reflected in AGI system, but also in affective computing research (*Picard, 1997*). Even the emotion recognition functions would need to analyze what the people experienced, facial expression recognition, and physiological change detection may be able to identify other's feelings, but without understanding why people experience certain emotion. If the system doesn't know what is the reason why certain emotion is triggered, the agent will not be able to make correct feedback.

People will not be happy about only one thing in their entire life, nor be sad for only one thing, so even for the same emotional state, it's not enough if the machine only has one way to deal with it.

Therefore, the importance of the appraisal process is to know why the individual is experiencing an emotional state, and to help understand the emotional state of other agents in the context of multi-agent interaction.

After understanding the importance of the appraisal model, the appraisal model should also be generic. What the appraisal model should evaluate is the information received by the system through all different sensors. The appraisal model is not only for a few events or a few kinds of events, but for all events, therefore, the appraisal model need to be generic.

Architecture Requirement 2: The level of emotional arousal should be determined by multiple elements.

It has been mentioned several times in this article that an event that triggers the emotion must be an event related to the agent's goal. Then, the primary measure of arousal intensity is the importance of the goal: the more important the goal is to the agent, the higher the arousal level will naturally be. Beside the importance of the goal, the influence level of the event on the goal also determines the arousal

level. Take fear for example, while fear is often defined as the emotion triggered when one's life and health are threatened, fear is more abstractly defined as the emotion experienced by an individual when an event that is contrary to one's goals is expected to occur (*Ortony et al.*, 1988; *Lazarus*, 1991).

Simply speaking, if B is a goal of the agent, an occurrence of event A may cause event B to occur in the opposite direction of the agent's desire. In this situation, if event A occurs, the agent will feel fear for the reason that B will be challenged. Therefore, when A occurs, the agent should be able to anticipate the gap between the upcoming event B and the expectations of goal B. The larger the gap between expectation of the goal and expectation of the anticipation, the higher the fear of the agent, and vice versa, the less fear.

The relationship between expectation of reality and goals' desire is the measure that initially determines the intensity of any emotion. For positive emotions, the smaller the difference is, the more the reality matches the desire and the higher the arousal level. Conversely, for negative emotions, the larger the difference is, the more the mismatch between the reality and the desire and the higher the arousal level.

In addition to the difference between desire values and reality, for some complex emotions, there are more factors could determine the arousal level. For example, nervousness or fear, in addition to the two factors mentioned earlier, should also be taken into account the difference between the expected time of the event and the current time. The smaller the difference, the faster something bad happens, and the higher the arousal, and vice versa.

Scherer's appraisal model considers nearly 15 factors for each emotion (*Scherer et al.*, 2001), although NARS did not choose Scherer's appraisal theory as the theoretical basis of the appraisal model (because NARS is not a cognitive framework based on numerical computation as the reasoning method). Lazarus's theory of emotion appraisal also refers to a multi-element assessment model. Therefore, emotional

arousal is by no means determined by a single factor. If the affective module is to be implemented in the AGI system, its cognitive architecture needs to provide similar appraisal criteria.

Architecture Requirement 3: An architecture that supports emotional modules should be capable of handling temporal information.

The importance of temporal information is not only to distinguish knowledge (time-free information) from events (time-dependent information), but more importantly, to distinguish events by tenses through the occurrence time of events.

The tenses of events are important for the appraisal of emotions, because some emotions can only be triggered by events that occur at a particular time. Worry or fear, for example, is triggered only because the agent anticipates something bad is about to happen, not necessarily because the event happened in the past.

On the contrary, sadness is only triggered for something that has happened or is happening (*Ortony et al.*, 1988; *Lazarus*, 1991), and not because of an anticipated bad event, although this statement is not absolute, because the agent may be very sure that something has happened by imagination, and therefore will be sad. However, if only the emotional states involved in the agent's interaction with the world are involved, then the tenses of events are fully capable of further refining the different types of emotions.

Therefore, an AGI system with an emotional module should have the ability to distinguish between events that occur at different times.

5.4 Summary

The requirements described in this chapter are all summarized in the current research process of NARS emotion module. These requirements can serve as the basic requirements of an emotion module of a cognitive framework, but are by no means limited to these requirements. Emotion is a complex cognitive process in human brain, and the activation of any emotion impacts the entire cognitive thinking process. With the continuous pursuit of emotional function in artificial intelligence, more and more complex emotions will have more requirements on the design, function and architecture of cognitive system, but at the same time, the cognitive framework will become more complete.

CHAPTER 6

CONCEPTUAL DESIGN OF EMOTION

MODEL IN NARS

6.1 Should Emotion be Innate or Learned in NARS?

There's a lot of discussion in psychology about "where emotions come from" (Izard, 1968; Barrett, 2017; Sauter et al., 2019; Damasio, 2011; Matsumoto and Willingham, 2009). The "how emotions come about" is different from the "how emotions are triggered" discussed in the previous chapters. When we talk about "how emotions are triggered", we're talking about existing emotional structures and their corresponding emotional functions in the human brain and NARS, but when we're talking about "how emotions are generated", we're talking about how emotion model are generated in cognitive frameworks that don't have emotional structures, just like many of the current cognitive frameworks that want to add emotional modules.

The outcome of the discussion on whether emotional modules are innate to cognitive system or developed through evolution will in part determine how emotional modules are implemented within the cognitive architecture of AI systems. Darwin insisted that emotions were a product of evolution, that people who survived natural selection developed emotions through different events and behaviors over the long course of human evolution (Darwin, 2013). This view has been espoused by many researchers (Barrett, 2017; Ekman, 2009), but at the same time, this argument is challenged by the requirement to maintain a strictly stable environment that suitable for evolution like which in the original evolution theory (Nesse, 1990).

Damasio’s view of emotions as unlearned, automatic, predictable, and stable behavior shows why they appear in natural selection and genetic instructions. Damasio also said that the basic emotional mechanisms in the normal brain are indeed highly similar, leading to a common basic preference for pain and pleasure across cultures (*Damasio, 2011*).

The problem of “where emotion mechanism comes from” in this paper is consistent with Damasio’s view that the emotional function itself is not learned. A large number of studies have proved that emotions emerge in infancy, and emotions play an important role in infants’ learning and expression (*Sullivan and Lewis, 2003; Thompson, 2001; Crockenberg and Leerkes, 2012*).

So even if the structures in the brain that process emotion are the result of thousands of years of human evolution, the evidence that babies are born with emotional functions makes it reasonable to assume we can implement similar functions in the cognitive architecture of AI systems in accordance with human emotional functions. There is no guarantee that the current AGI systems can be replicated through a process like human evolution. Even if the system is capable of processing complex information, is it fair to use human survival experience as the emotional learning content of AGI system? AGI system is not designed to completely simulate the thinking process of the human brain. Differences in culture and education lead to differences in emotional triggering conditions between people. It is inevitable that the perspective and way artificial intelligence systems look at the world is different from that of human beings, so it is inappropriate to use human survival experience as the learning content to make artificial intelligence system acquire emotions.

Emotions in NARS are partially learned, even though the mechanism is innate. For humans, almost all concepts are learned (*Drescher, 1986*), so even if the mechanism that triggers an emotion is not learned, the situation that triggers an emotion is. The stimuli that trigger emotions are influenced by different factors such as cultural

background, personal experience and education. As a result, people from different cultures may have the same emotions, but they may express different emotional states for the same thing. An experienced hunter, for example, would be excited by the prospect of a new prey when he saw a wolf, but an ordinary person, who had no experience in hunting, would be frightened by the threat to his life. Similarly, coping behavior after emotional trigger is learned from experience. An experienced hunter may quickly dodge when suddenly attacked, but an ordinary person may stay put and get hurt. The main role of emotion is not to directly trigger behavior, but to generate corresponding motivation through internal resource regulation to promote the generation of certain behavior. If the behavior is not learned, even if the corresponding emotion is triggered, there will be no corresponding response behavior.

The implementation of emotion modules in the cognitive framework of NARS fully follows the results discussed above, that is, emotion function is innate, but cognitive appraisal, and coping are entirely learned by NARS. Emotion in NARS is divided into implicit expression and explicit expression. The emotions of implicit expression are hidden in the functions of NARS, while the emotions of explicit expression have specific appraisal criteria.

A Lesson Learned From a Previous Design

In a previous study (*Li et al.*, 2018), the author used a method similar to evolutionary emotion theory to make NARS have emotions through learning. Although this method is not an appropriate method at present, it is an attempt to see the possibility of realizing emotions in the general artificial intelligence system. At the same time, it is also found through this method that learning with mental operators in NARS and making NARS have emotions cannot solve the emotional problems well. The following section will review the previous solutions and summarize the problems of the previous methods.

Here we will discuss two examples from previous studies, which can be found in Appendix A. In the first example of happiness, when the desire of the system is satisfied, then the system triggers happiness. The emotion itself does not adjust the internal resources of the system, and no corresponding behavior is activated. Therefore, at the present stage, learning with mental operator cannot trigger the corresponding adjustment of internal resources, which is not consistent with the function of emotion in psychology. In the statement of learning:

```
// The meaning of this statement is if "SELF" desires something
// to happen and believes the thing is already happened,
// then "SELF" feels happy
Input: <(&&, (^want, {SELF}, #1, TRUE), (^believe, {SELF},
#1, TRUE)) =|> (^feel, {SELF}, happy)>.
```

This statement only triggers a feeling, not an action. If the system continues to learn, where a behavior can be triggered when the system is happy, the emotion becomes a direct precondition for the behavior, not a diversion of attention through resource adjustment, but a direct precondition for the emotion through implication. Obviously, this is not the same mechanism as human emotion. In the second example of fear, when the system is afraid and motivation to escape is generated through the “want” operator, the priority of the escape is increased by the “want” operator, but in the fundamental sense, the priority of the escape behavior is not directly related to the threatened goal. Even though fear is an emotion triggered by the event that the agent is about to be injured. In this sense, the agent has no way of knowing why it is afraid. Without knowing why it is afraid, how can the system choose the right behavior among the many responses to fear based on the current situation?

When only searching for the direct implication between environmental status and behavior, then emotion is unnecessary. If the implication could ensure that behavior is triggered under certain circumstances, does this prove that AI systems don't

need emotions? However, simply looking for an implication between environmental change and behavior would require an exhaustive list of various corresponding relationships, which would run counter to the basic assumptions of AIKR in the first place. In addition, if, as in the case of fear, a motivation to flee is generated, and the system mechanically triggers the motivation and behavior under certain pre-defined circumstances, does the system lose its autonomy?

Therefore, the previous development of NARS is not suitable for making NARS have emotion nor proving the necessity of emotion by the method of evolutionary emotion theory.

6.2 Architecture of Emotion Module in NARS

Emotion module that Processes External Stimuli

Starting with Dr. Wang's doctoral dissertation in 1995 (*Wang, 1995*), NARS has gone through a series of editions for nearly 30 years. The design of NARS is not based on any existing human psychological models, but based on a integration of different theories from psychology, philosophy, and related disciplines. Other AGI systems such as LIDA are designed based on global workspace theory (*Franklin et al., 2014*) and MicroPsi (*Bach et al., 2019*) is based on Psi cognitive theory. The architecture design of NARS was adjusted several times throughout the development process. OpenNARS 3.0.4, the last Java release before this paper was written, added temporal inference and the ability to deal with goals on top of OpenNARS 1.5.8, and already has simple emotional functions, like satisfaction and busyness, as well as some complex emotions, such as anxiety, fear, sadness. Even so, the previous version of the system's emotion-processing framework had many problems. The first problem is that, in the previous cognitive framework, emotion is merely a plug-in to the main program, not a part of the main cognitive framework.

Design with emotion module as a plug-in has several disadvantages. Firstly, the emotion module is linked to the main program as a plug-in. When an event enters NARS, it does not interact with the emotion module immediately; the emotion module needs to traverse all the events in the current event list to determine whether any event meets some appraisal criteria of the emotion. However, the events in the event list are not guaranteed to be current, so an emotion that should promote quick reaction may have lag, and traversing the event list cannot be guaranteed to avoid repetition. Also, the inefficiency of traversing the list is self-evident when a large number of events continually flood in.

The second drawback is that the emotions triggered by the cognitive resources adjustment in NARS “main memory”, and the result of the cognitive resources adjustment, will take effect on the next working cycle, rather than the current work cycle. This is similar to the previous disadvantages: handling cognitive resource adjustment in this way for some rapid response motivated by emotion can be fatal, for example fear. When the agent has realized that the danger is approaching, but the fear caused by the danger does not affect the current reasoning cycle, the agent still continues the original thought process in the current reasoning cycle, instead of reacting quickly to avoid the danger due to the fear.

According to Damasio’s description about the process of Emotion-Feeling cycle, emotions, starting with the potential stimulus that triggers emotions, stimulates awareness and appraisal then spreads to other areas of the brain and body, forms the emotional state, and finally, back to the brain, forms the feeling, in a different brain region than when it was initially primed (*Damasio, 2011*). From this description, it is not difficult to see that emotion starts from stimulus, goes through the adjustment of cognitive resources to form feelings, and this is the same cognitive cycle, rather than acting on the next cognitive cycle. NARS, therefore, requires a reasoning framework which does some basic reasoning before the reasoning in the main memory, this frame-

work is not a separate architecture from the main architecture, but an architecture to make sure that, when an external stimulus is realized by NARS, corresponding emotion can be triggered in this “pre-conscious” framework, and cognitive resources brought by the emotional adjustment will take effect in the current cycle of reasoning. For example, when the risk of sudden is detected by NARS, the goal of “keep health” is threatened and fear emotion is triggered, the emotion will make sure the goal “keep healthy” is activated before it enters the inference with consciousness. A big increase in goal’s priority to ensure that the corresponding concept of the goal will be selected as soon as possible, and the coping will be execute by NARS based on the goal and the current situation.

Emotion Module that Processes Internal Stimuli

In addition to current events that can trigger emotions, “cognitive signals” can also be used as stimuli that trigger emotions. Such “cognitive signals” can be memories of past events, or conjured up scenes by imagination, which may trigger the corresponding emotional experience. This also explains why people experience emotions when listening to music or watching a video. Similarly, when people hear about another person’s experience, they also experience that person’s emotional state to some extent, but with different intensity and response than the original emotional experienter.

Emotions triggered by external events will directly affect the resource allocation of current working cycle in main memory. The emotion module which handles internal event processes the information delivered from the reasoning in the main memory. The information delivered from main memory may include memories recall, reasoning result, and imagination. Technically speaking, the information that reaches the “post-conscious” region comes from the inference of the current working cycle in the main memory area. The emotional experience will occur immediately, but the cognitive

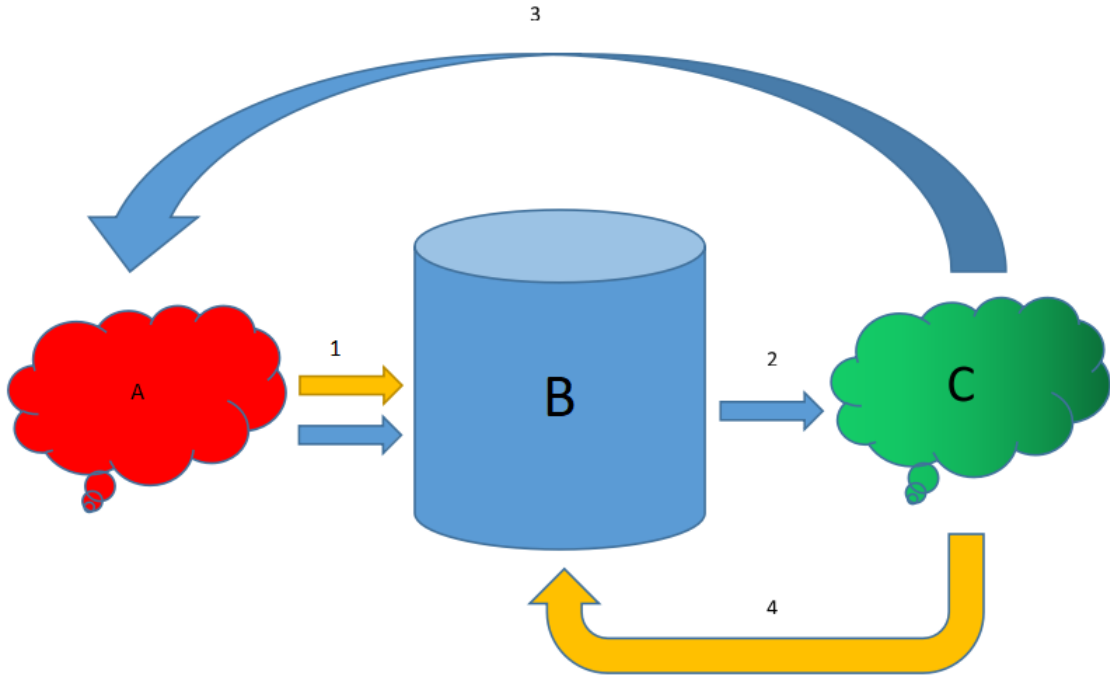


Figure 6.1: Task flow and Emotional flow within the target framework

resource regulation corresponding to the emotion will be reflected in the next working cycle.

Figure 6.1 simply describes the relationship between the emotion module of NARS and the main memory. The figure is not the complete structure of NARS, which will be introduced in detail in the next chapter. In the figure, module A is the emotion module dealing with external events, module B is the main memory, and module C is the emotion module dealing with internal cognitive stimuli. The blue arrows are task flows and the yellow arrows are emotional flows.

When task T1 passes through module A, if emotion E1 is triggered, it will immediately adjust the cognitive resources in main memory (B). Whether or not the task triggers emotions, task T1 is passed from A to B for processing. In addition to the processing of task T1, B will also extract and process the concept in the existing memory. Whether processing the new task or processing an existing concept, the result will be passed to C as a task. The derived task in B is the result of the reasoning

process in the main memory, which can be explained from the perspective of reality, the result of the reasoning in the main memory (which can be the memory of the past), reasoning based on the present situation, or the reasoning based on reality or made-up imagination. So let's say that the result of B is a task T2, and whatever T2 is, T2 is going to be sent to module C, and then it's going to be processed emotionally in response to internal cognitive stimuli. Whether or not T2 triggers emotion, T2 will be sent back to A for processing in a new cycle of work. However, if emotion E2 is triggered by T2 in C, then E2 will go back and reallocate the cognitive resources in main memory B.

In each working cycle, B will be adjusted by triggered emotion from A and C. If no emotion was triggered from A and C, then B will be doing the default reasoning process. If any emotions are triggered from A or C, B will still be doing the normal reasoning process, but the difference is that the structure of the internal resource competition has been affected by emotion, therefore, there is the possibility that the concept chosen by B for reasoning in the current working cycle will change.

In conclusion, the new emotion architecture of NARS has the ability to evaluate external and internal stimuli. When the potential stimulus meets the appraisal criteria, the emotion module of the current task will generate emotion and adjust the cognitive resources in the main memory, and the adjustment results will be applied to the current or the next working cycle. This workflow is consistent with Damasio's description of the emotional process (*Damasio, 2011*), in which any emotion starts in one area of the brain, expands the emotional processing to other areas of the brain, and eventually the processing returns to the brain and forms a circuit. The destination is a different part of the brain from the area where the emotion is activated. In the emotional architecture of NARS, emotions can be initiated in the external stimulus processing module or the internal stimulus processing module, but feelings will eventually return to the main memory.

6.3 How many locks can a “key” open?

The previous section describes the design of the emotional architecture of NARS. To put it simply, the emotional system in NARS is not completely “learned”. The structure for dealing with emotions is predetermined, but the conditions which trigger emotions and the coping are learned from experience. After that, two cognitive structures containing emotional modules should appear before and after the main memory, respectively, to process sensory stimuli and stimuli from the internal mind, respectively. The structure of the emotion factory has been introduced, now let’s introduce the assembly line of the emotion factory.

NARS is a general artificial intelligence system. The modification, addition or even deletion of any module cannot affect its generality. The emotion module should also have generality.

Hypothesis 1: The generality of emotion in AGI system is manifested in two aspects. First, the emotional trigger conditions should correspond to all events in real life, not to specific events. Second, emotions do not work for any particular function or task, but open to all tasks and goals which related to agency.

The ultimate function of emotion is to maintain individual homeostasis. Homeostasis is the stability of the sum of all the needs of an individual. When random events disrupt, or threaten to disrupt, this stable state, emotions appear in an attempt to restore the original stable state. (*Damasio, 2011*).

Therefore, as a general intelligence system, AGI emotion module should also be oriented to any kind of event, even if it unexpected or not previously encountered. Therefore, for each different emotion, there should be an abstract appraisal model.

When the abstract meaning of an event is consistent with the corresponding appraisal model, the corresponding emotion will occur.

Valence of Emotion

The so-called abstract meaning is the relationship between the event itself and the agent. According to Lazarus' cognitive emotion theory, only an event related to the goal of the agent can trigger emotion, while an event unrelated to the target will not trigger the emotion. Goal relevance is the first criterion of emotion appraisal. When an event has nothing to do with the goal of the agent, it will be rejected by any emotion appraisal model and will not be further evaluated.

In addition to the base emotion "satisfaction" in NARS, previous studies have tried to add more complex emotions into NARS (*Li et al.*, 2018). Goal relevance identifies whether an event can trigger an emotion, and after identifying the goal relevance, the next step is to identify the valence state of the emotion. The valence of emotion can determine whether an emotion is positive or negative to the agent, which is the second judgment after goal relevance. Without subdividing different emotions, the valence state of emotion is helpful to understand the effect of valence and arousal on the cognitive processing of human brain (*Kauschke et al.*, 2019; *Shuman et al.*, 2013; *Citron et al.*, 2014).

Lazarus uses goal congruence to express valence. After determining whether the event is related to the goal of the agent, goal congruence will be determined based on whether the event is consistent with the goal. If the event is consistent with what agent desires, positive emotion will be triggered, otherwise, negative emotion will be triggered.

In NARS, goal congruence can be evaluated by comparing the desired value of the goal with the truth value of the event. Generally speaking, for the desire value or the truth value, if the corresponding Narsese expectation is 0.5, then it is neutral. For a

Table 6.1: The relation among desire-value, truth-value, and the valence

Desire value	Truth value	Valence
want	have	positive
want	not have	negative
not want	have	negative
not want	not have	positive

Table 6.2: XNOR Gate Truth Table

Input 1	Input 2	Output
0	0	1
0	1	0
1	0	0
1	1	1

goal, it means that NARS itself has no obvious preference for the goal. However, if the expectation is greater than 0.5, it means that NARS wants the goal event to happen, while less than 0.5 means that NARS does not want the goal event to happen. The level of desire depends on the difference between the value of desire and 0.5. The bigger the difference is, the stronger the desire will be, and the smaller the difference, the weaker the desire will be.

In (*Li et al.*, 2018), the appraisal of goal congruence in NARS has already discussed in detail, and found that the combination of desire value and truth value and corresponding valences of emotions matches XNOR gate logic completely.

Table 6.1 describes the Goal desire values, event truth values, and the corresponding valence of the emotion. “want” represents NARS desire some event to happen, and “not want” indicates that NARS desire some event not to happen. Similar representations to Truth value. Table 6.2 is the truth table for the XNOR gate. The determination of valence corresponds to the relationship between desire value and truth value in the abstract sense, and such relationship is universal, not for a certain kind of goal or event.

In this sense, the emotional mechanism of NARS still meets the requirements of a general system. But emotions are more than just positive and negative, and further judgment requires a deeper level of analysis. Lazarus calls it ego-involvement, a mechanism for determining the relationship of events to the agent itself, and the analysis of this step can directly determine which unique emotions will be triggered in the system. Part of Lazarus's description on ego-involvement is too specific to apply to NARS at this stage, because many relationships are too specific, such as an event related to the health of the individual, or self-esteem. Frankly speaking, NARS at this stage is not capable of understanding what dignity is, so Lazarus' elaboration of ego-involvement cannot provide much inspiration for the emotional module of NARS architecture at this stage.

Appraisal Model for Specific Emotions

In addition to Lazarus, many psychologists who studied emotions also developed a number of emotional appraisal models. Denham (*Denham et al.*, 2002) also designed an integrative model of three components (i.e. desire, state, and belief of certainty underline the cognitive process) for a child's experience of different emotions. The **Prototype Approach** describes the correlation between general types of events and specific emotions; each emotion is linked to common situations that cause it. For example, pleasurable stimuli or getting or doing something desired causes happiness. Anticipated harm or unfamiliar situations may cause fear, etc. Instead of encompassing emotional themes, the **Event Structure Approach** focuses on capturing the processes by which children come to experience different emotions. A child may experience fear if she realizes that she is very unlikely to maintain a desired state. In contrast, a child may experience anger if he realizes that some external conditions prevent him from achieving a desired state or avoiding an undesired state. The last approach is called **Desire-Belief Approach**, and it describes how emotions may

Table 6.3: Integrative model of Happiness, Sadness, Anger and Fear

Desire	want	want	want	not want
State	have	not have	not have	have
Belief of certainty	yes	never	can reinstate	likely
Emotion	Happiness	Sadness	Anger	Fear

result from the consistency or discrepancy between one's desires or beliefs and the reality. A child who desires a gift feels happy if he actually gets one; in contrast, a child who believes Mom is sleeping in the bedroom may feel surprised when she finds nobody there. Based on these three components, Denham proposed an integrative model that encompasses both the process and the content of a child's reasoning that leads to different emotions. Table 6.3 shows the model for Happiness, Sadness, Anger, and Fear.

Ortony, Clore and Collins introduced a clear and easily implemented appraisal model in a computer system (*Ortony et al.*, 1988). This model was also used by the authors in their own computer program to verify the effectiveness of the model, even though the program was not intended to implement AI emotions, as the authors thought such efforts would be futile. OCC model emphasizes that the cognitive interpretation of the event is the key to trigger emotion. OCC model is similar to Lazarus' cognitive emotion theory model, both of which indicate that whether an event triggers an emotion depends entirely on the agent's cognitive interpretation of the event. This explanation is not necessarily based entirely on the event itself, but may also be based on the consequences of the event. In NARS' emotion model implementation, this problem has been solved: after the occurrence of an event is detected by the system, the event itself will be sent to the emotional appraisal module related to past events, such as sadness, happiness, surprise; at the same time, the consequences of events associated with future events will also enter the emotional

appraisal module to trigger, for example, fear and anxiety.

Hypothesis 2: The appraisal model of emotion in Artificial General Intelligence systems should not only consider whether the current event is able to trigger emotions, but also consider whether the consequences of the event can also trigger emotions

With a brief look at the work on affective cognitive assessment, let's go back to the original question in this section. How many locks can a key open? If we think of an event as a key and different emotion assessment models as different locks, when an event occurs, only one emotion will be triggered, or is it possible to have multiple emotions at the same time? Denham's model and the OCC Emotion Model do not appear to support multi-emotion concurrency on the surface, because there is no overlap between the appraisal models. So is it really true in this metaphor that "a key only opens one lock? "

Consider a simple virtual scenario where Tom's mother gives him 10 dollars to buy a bag of rice at a nearby store, and Tom stops there for 5 minutes because he is attracted by a roadside show on the way to the store. But when Tom chose the rice and was ready to pay, Tom found the money was missing. That's the end of the story. What possible emotions are triggered when Tom discovers that his money is missing?

First of all, losing money can trigger different emotions, such as sadness, because the money is lost, anger, if Tom thinks the money was stolen, regret, because Tom watched the performance and lost the money. These three emotions are actually caused by the failure to achieve the goal of buying rice due to the loss of money. In addition to these three emotions, Tom may also feel afraid or anxious because he think may be scolded by his mother for losing money and failing to complete the task

assigned by his mother. Therefore, the occurrence of an event is likely to trigger many emotions, perhaps these emotions are not directly caused by the event, but may be caused by past and anticipated future events or thoughts associated with the event.

Hypothesis 3: The emotion module of a AGI system needs to have the ability to trigger all emotions by events that meet the appraisal model, rather than just triggering one emotion when multiple emotions meet the criteria.

The appraisal model of emotion module in NARS is an appraisal model which is suitable for NARS' logic and control system based on the synthesis of several appraisal models of emotion in cognitive emotion theory. Regardless of the emotion, according to Lazarus' theory of appraisal, the goal-relevance of the event is first judged, and if the event is unrelated to the objective, the emotion appraisal model is skipped, otherwise, the appraisal will be performed.

Based on the current expressive power of NARS, as well as the practicality of emotion, seven emotions have been realized in NARS. The seven emotions are hope, satisfaction, disappointment, fear, relief, anxiety and sadness. The Appraisal model of these seven emotions is as follows:

- **Hope:**

- If the system wants E to happen, and the system anticipates E is going to happen.
- If the system does not want E to happen, and the system anticipates E is not going to happen.

- **Satisfaction:**

- If the system wants E to happen, and E has happened.

- If the system does not want E to happen, and E is confirmed to not have happened.

- **Disappointment:**

- If the system desires E to happen and anticipates E to happen, but E does not happen.
- If the system desires E not to happen and anticipates E won't happen, but E happened.

- **Fear:**

- If the system wants E to happen, but the system anticipates E is not going to happen..
- If the system does not want E to happen, but the system anticipates E is going to happen.

- **Relief:**

- If an event is feared by the system, but the result is different from what the system anticipated.

- **Sadness:**

- If the system wants E to happen, but E didn't happen, and the result is irreversible.
- If the system doesn't want E to happen, but E happened, and the result is irreversible.

- **Anxiety:**

- If the system wants E to happen, but system is not sure if E is going to happen.

- If the system doesn't want E to happen, but system is not sure if E is going to happen.

Here, take fear as an example to illustrate the whole process of emotion in NARS. Although fear is classified as a negative emotion in terms of emotional experience, the attention diverted by fear and the positive attitude generated for survival are very important to ensure the individual's survival and that when a goal is threatened, the system can divert attention through fear and respond quickly.

According to the above appraisal model, fear is a response to a potential conflict with the goal. If explained in terms of the control process of NARS, first of all, NARS has learned or discovered the temporal sequence relationship between event A and event B through temporal induction, $A \not\Rightarrow B$. This relationship can be interpreted as "if A happens, then B will happen", or, of course, it may not happen, if the corresponding truth expectation of this relationship is less than 0.5. Now assume that the Narsese expectation is greater than 0.5, that is, when A happens, B will happen a certain time later, so when the system detects A happens, it will generate an expectation that B will happen at a certain point in time. But if the system doesn't desire B to happen, then at this point, the expectation conflicts with the goal, and fear is triggered.

The above example combined with the appraisal model of fear can exactly confirm the relevant discussions in Chapter 7 and Chapter 8. First of all, the emotion module system should have the concept of goal. Without the concept of goal, the triggered emotion is meaningless. Second, the concept of time is important; emotional satisfaction and sadness are for things that have happened in the past, but hope, fear, and even anxiety, are for things that have not happened yet. Therefore, without the concept of time, the system cannot distinguish between emotions.

6.4 The Functions of Emotion

As mentioned in the previous section, one of the important functions of emotion is to maintain the “homeostasis”. (*Damasio*, 2011).

“Homeostasis” is defined differently within different bounds. When describing the “homeostasis” of an individual’s body, it refers to the stable state of vital signs and internal functions of the body. When the stable state of the body changes, emotions will encourage the agent to correct the unstable state. In the same way, homeostasis can extend beyond the individual, such as the stable state of a relationship between two people, the stable state within a family, and the stable state within a social group. Regardless of the type of object, there is an implicit (unaware of the individual, inadvertently maintaining homeostasis) or explicit (aware of the individual, intentionally maintaining homeostasis) goal in the individual’s mind, which is to maintain “homeostasis.”

Is Emotion Same With Motivation?

If the most significant function of emotion is to maintain “homeostasis”, then when homeostasis is abnormal, emotion will be aroused to maintain homeostasis, then emotion can be regarded as motivation? Also, as discussed earlier, emotion is triggered by an event associated with a goal. Could emotion be motivation?

Before we discuss the relationship between emotion and motivation, we must first know what motivation is. Motivation refers to the internal mechanisms that trigger proximal behavior based on its direction. In other words, it promotes some behaviors while inhibiting others (*Kleinginna and Kleinginna*, 1981).

The reason that emotions are often seen as motivation is that some emotions are thought to trigger behavior. For example, people want to run away when they are afraid, or they want to destroy when they are angry. There are a lot of similarities between emotion and motivation, but there are also a lot of differences, which means

that emotion and motivation can't be equated. According to Roseman's analysis of motivation and emotion (*Roseman, 2011*), the similarity between motivation and emotion lies in that both of them belong to internal states or processes, and both can lead to goal-oriented actions, while the difference lies in that motivation needs to be activated by specific conditions, while emotion can be generated by accidental events applicable to any motivation. For example, the motivation to eat is generated by hunger, the motivation to drink is generated by thirst, but happiness can be triggered by either goal being satisfied.

Therefore, emotion is not entirely a motivation. If the difference is explained in the framework of NARS, motivation is the desire of NARS to take action to achieve a specific goal, while emotion is the feeling when real events or thoughts of NARS occur and they have a certain connection with a goal. Emotional feelings contain a degree of motivation that motivates NARS to take actions to maintain or prevent the impact of the event on the goal.

Emotion is like motivation and would contribute to different behaviors. Roseman points out that emotional behavior is goal-oriented in some cases and is accompanied by instrumental behavior aimed at achieving specific emotional goals (*Roseman, 2011*). Lazarus, on the other hand, points out that subjective influences, physiological changes, and predisposition to certain behaviors during the process of emotional production, can distract attention from the activity in progress (*Lazarus, 1991*).

Thus, emotions do not directly generate behavior, but rather adjust the agent's attention through the reallocation of internal resources. Different emotions have different ways of adjustment. To put it simply, negative emotions narrow the attention of the agent and make the attention focus on the events that trigger the negative emotions, which helps the agent to solve the problems that hinder the goal. For example, when the agent feels fear, fear will motivate the individual to focus on the thing or event that causes the individual to feel fear and try to solve the problem

through experience.

In NARS, different emotions will be applied to adjust internal resources such as priority and durability of corresponding concept, and adjust the attention of NARS. It will also give NARS ability to deal with real-time task, even when the task is unexpected, and NARS can still do planning for the current situation on its own, or adjust the relationship between NARS and the environment as appropriate.

Take fear for example. When humans experience fear, it is natural to focus on the event that triggers the fear emotion and begin to focus on finding a way out of the situation. Therefore, fear can help agent to focus on a potential problem that may have a negative impact on the agent and urge the individual to solve the problem as soon as possible. While the method to solve the problem originates from the agent's experience, fear can prompt the agent to choose a solution that resolves difficulties based on the current situation (e.g. remove the threat of the target). Therefore, the emotion does not directly correspond to the behavior, but finding and selecting the best solution by transferring attention based on the current situation and past experience.

6.5 Summary

This chapter introduced the theoretical foundation and implementation framework of the emotion model in NARS. Firstly, the emotion module of NARS is innate. In other words, the framework of the emotion module, the cognitive appraisal model, and the resource allocation function will appear in the system as part of the program code. This approach does not make NARS itself lose the generality, and emotion modules are also general. In the process of cognitive appraisal, the appraisal model will be oriented to any event in real life. In the future application process of NARS, there is no need to set special tasks and events for the emotion module. The appraisal model evaluates the abstract meaning of the event for the goal of NARS, without caring

about the actual significance of the event.

The “dual emotion module” is located before and after the main memory, and acts on real events perceived by NARS, and the results after the NARS “brain storm”. The results of the emotion effects are fed back into the main memory and begin to take effect in the most recent work cycle.

In the next chapter we will introduce a new NARS architecture that builds on OpenNARS 1.5.8, adds temporal reasoning, targets, and replaces the old buffer architecture with a new buffer architecture. The new buffer architecture perfectly supports the emotional model described in this chapter.

CHAPTER 7

EMOTION MODEL IN NARS

This chapter will introduce the concrete implementation of the emotion model in NARS, as well as the upgrades to the new architecture of NARS to support the appraisal model in the emotion module, the realization of the emotion function, and the testing of practical examples.

7.1 Structure of Buffers

The structure of the Overall Experience Buffer and Internal Experience Buffer are similar in general, although there are many functional differences. The buffer is equivalent to short-term memory in the human brain. Both the Overall Experience Buffer and the Internal Experience Buffer screen the information entered into the buffer according to importance (priority) and time. The buffer sends the most important information from to the next processing module based on priority, and at the same time eliminates unimportant and outdated information to free up resources and processing space for new information.

In addition to screening information, the more important role of buffer is to preprocess the information entered into the buffer; this preprocessing includes sorting, ignoring, temporal induction, anticipation, and cognitive appraisal and expression of emotion.

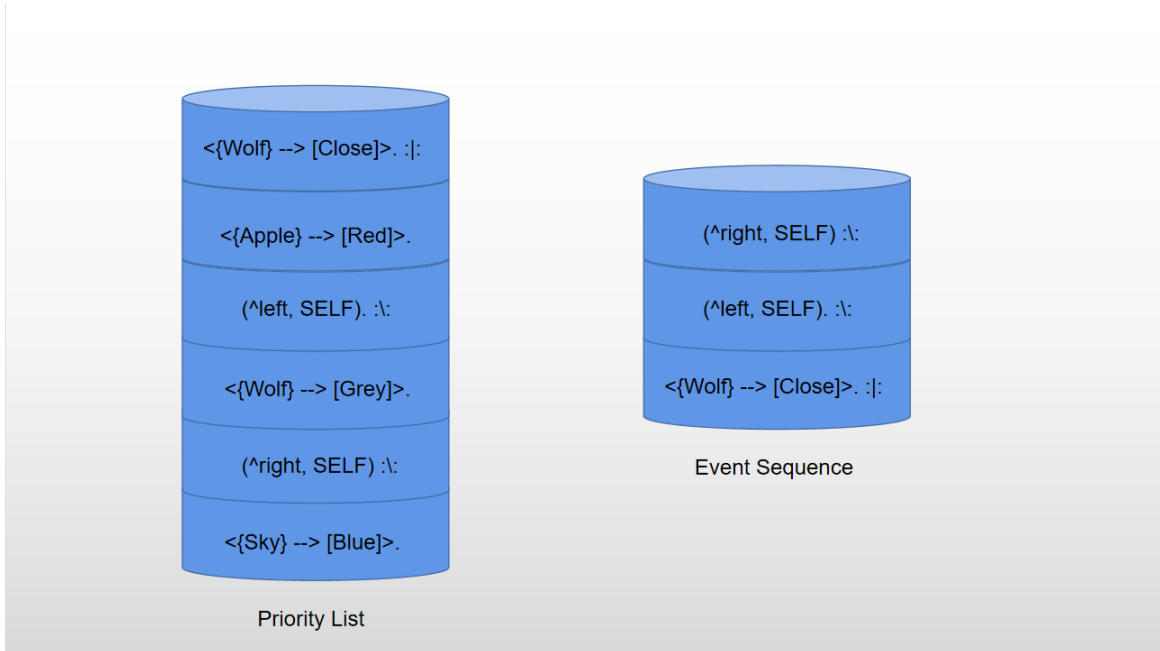


Figure 7.1: Example state of a buffer

Storage structure

Each buffer has two storage structures. One is for storing all information that enters the buffer, whether events (with a time attribute) or knowledge (without a time attribute). The storage is sorted by priority, aiming to first send the most important information to main memory, or to the next storage carrier. The other storage structure is the event sequence, sorted by occurrence time of event. This sequence and the priority queue may have overlap, because an event can appear in both sequences at the same time, but a task without a time attribute will not appear in the event sequence. The main purpose of the event sequence is to do temporal induction and find the temporal relationship between two or more events.

As shown in Figure 7.1, the priority list can hold any information that goes into the buffer. It can be events such as “< {*Wolf*} → [*Close*] > . : | :”, it can be goals such as “< {*SELF*} → [*Good*] > !”, it can be operations such as “(*op(left), SELF*).”, but the event sequence can only hold events, and non-event information will not go

into the event sequence. Neither list is infinite, just as human short-term memory is not infinite, it's capacity is limited and the priorities of items in the list decay over time (*Cowan, 2008*). Capacity is flexible and can be configured according to the application, but it is never assumed to be infinite. In fact, setting the buffer capacity to infinity does not make any practical sense, because when NARS is applied to a real scenario, events will continue to enter the buffer, and things that are noticed by the system but are not important will only remain in the buffer storage resources and never be selected by the system.

Task Flow

As shown in Figure 4.2, NARS receives information from the outside world through different sensors (channels) and temporarily puts the perceived information into the Overall Experience buffer. At the beginning of each working cycle, the Overall Experience buffer will also take a task from the Internal Experience buffer and place it in the Overall Experience buffer. Based on the previous discussion of the NARS control mechanism, both events with a time attribute and knowledge without a time attribute are treated as a task by NARS. Each task has a budget value, and in both the Overall Experience buffer and the Internal Experience buffer, all tasks in the priority list are sorted by priority.

According to selective attention theory, intelligent behavior requires selecting and focusing on specific inputs for further processing, while suppressing irrelevant or distracting information (*Hanania and Smith, 2010; Stevens and Bavelier, 2012*). The source of information in the Internal Experience buffer is only the inference results of the main memory, while the Overall Experience buffer contains the inference results of external (sensory) information and the Internal reasoning result. Therefore, when a task tries to enter any of the buffers, the priority of the task is used to determine whether to insert the task into the buffer. Buffers do not have pre-set priority thresh-

olds, so the criteria for determining whether a task can be inserted into the buffer varies from case to case.

If the priority list within the buffer is not full, the task can be inserted into the priority list regardless of its priority, and sorted according to the priority. However, if the buffer is already full, then one of two possibilities will happen: 1). If the new task has a higher priority than the lowest priority task in the priority list, then delete the lowest priority task in the priority list and insert the new task into the corresponding position in the priority list. 2). If the priority of the new task is lower than the lowest priority task in the list, the new task cannot be inserted into the priority queue. This design is consistent with the selective attention theory. Even if a task with low priority is perceived by the sensor, it will be ignored by the system and no further processing will be conducted. This process occurs for each buffer.

Each time a task tries to enter the buffer, the task first attempts to be inserted into the priority list, and if the task is an event, then after successful insertion into the priority list, the event is further inserted into the event sequence according to its time of occurrence. If the task is an input with no time attribute, no attempt is made to insert into the sequence of events. Finally, if the task is an event but is too low in priority to be inserted into the priority list, no further attempt is made to insert into the event sequence.

According to Hammer (described in (*Hammer et al.*, 2016)), no matter whether a task is an event or knowledge, after it is inserted into the buffer, the task can only stay for a certain amount of time in the buffer. If a task remains in the buffer and is never selected for processing by the next module, the task will be deleted from the buffer directly, its significance is forgotten by the system or ignored.

This operation ensures that the relatively important information is always kept in the buffer, and the unimportant information is excluded by the buffer. When more important information tries to be added to the buffer, the unimportant information

currently kept in the buffer will be ignored or forgotten by the system. Only when the system is idle, the unimportant information may have more chances to be processed; however when the system is busy, only the most important information will win the resource competition and get the opportunity for further processing.

Temporal Induction

According to AIKR, inference in NARS' master memory is not arbitrary, but must occur between semantically related statements. However, in the buffer, NARS allows temporal induction for non-semantically related, but rather time-related events (*Hammer et al.*, 2016). In this sense, events with close occurrence time can generate the corresponding implication relationship and event sequence through time induction. Restricted by Duration in the buffer, two events with time interval longer than duration will not trigger temporal induction because the earlier event will be deleted by the buffer before the later event enters the buffer.

In the current design, temporal induction only generates two kinds of relationships: forward implication ($E_1 \Rightarrow E_2$) and forward event sequence ($\&/, E_1, E_2, \dots$). Temporal induction happens when an event is trying to be inserted into the buffer, rather than when the events are selected into the main memory, because when testing with the environment which will be introduced later, when a large number of events are constantly poured into the buffer, if temporal induction happens when the event is selected from the buffer and to be processed in the main memory, many events may be blocked in the buffer without getting a chance to be processed, the result is that the event will eventually be deleted since it stay in the buffer for too long and get deleted by the duration limitation.

Figure 7.2 briefly shows the process of Temporal Induction. For convenience, we use a, b, and c to represent the three events. OT is the Occurrence Time of the event. Initially, only event *a* was in the event sequence list. By stage 2, event *b*

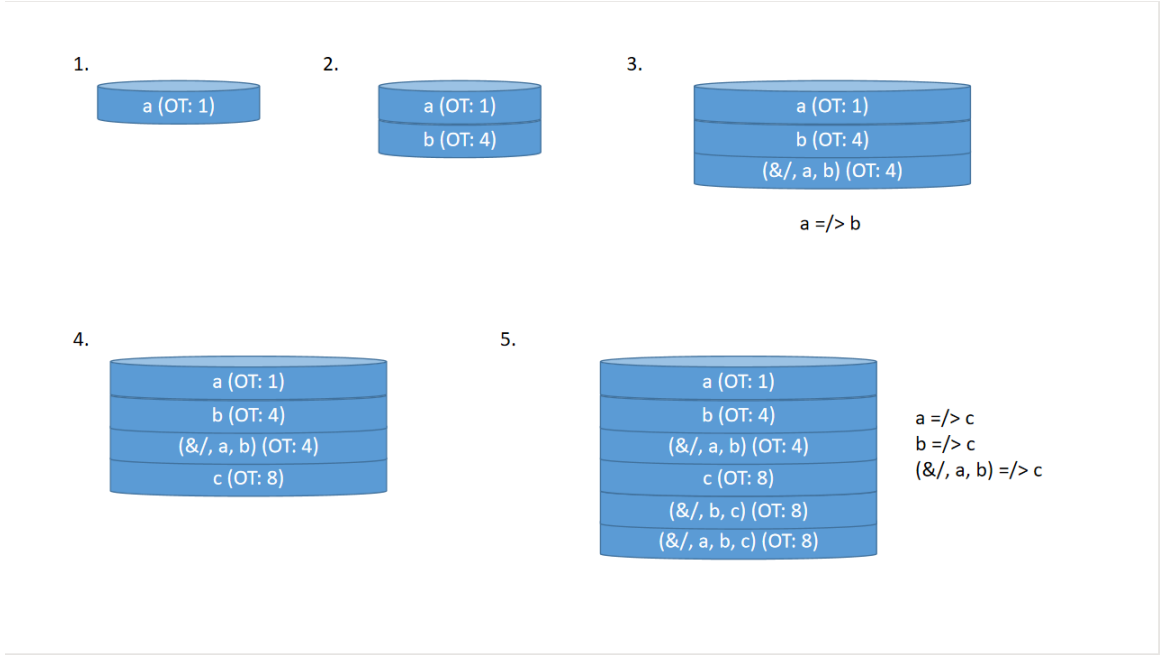


Figure 7.2: Temporal Induction in the Event Sequence of Buffer

was perceived and successfully inserted into the buffer. Because b occurs later than event a , b is inserted after event a . After successful insertion, since there is no third event between event a and event b , event a and event b undergo time induction, and generate implication “ $a \Rightarrow b$ ” and event sequence $(\&/, a, b)$ at the same time. Because implication describes a process, that is, “if a happens then b happens”, the implication relation is not regarded as an event, but this knowledge is directly sent to the main memory. However, what event sequence “ $(\&/, a, b)$ ” describes is the sequence of occurrence of a series of events, which does not contain the “if-then” causality in its semantics. It is just a description of a compound event, that is, b happened after a happened. Thus, the sequence of events is still a compound event with a temporal property. For compound event “ $(\&/, a, b)$ ”, after b occurs, the whole compound event is perceived by the system and the sequence of event a and event b is concluded through temporal induction. Therefore, the occurrence time of a compound event is always the occurrence time of the last event in the compound

event. Therefore, in Stage 3, the compound event is inserted after event *b*. In a practical sense, event *b* and compound events are not chronologically sequential.

When time comes to stage 4, a new event *c* is perceived by the system and successfully inserted into the buffer, and a more complex round of time induction will occur in the event sequence. First of all, event *c* will be implied by any event that occurs before event *c*, because in the buffer, the system will not consider the semantic correlation between events, so any event that occurs within the same duration will be considered potentially causal. Whether the causality is correct or not is not taken into account in the buffer. The specific discussion on this issue will be carried out in the following content. Secondly, event *c* only forms an event sequence with events that strictly occur before *c*. According to the discussion just now, compound events describe a series of events with time properties and occurrence sequence. So, event *c* cannot form a compound event with event *a* alone because *b* occurs between *a* and *c*.

However, a compound event containing three events *a*, *b*, and *c* is allowed to be generated because the compound event describes the entire sequence of events. This is why the time of occurrence of the compound event is marked as the time of occurrence of the last event. In addition to the temporal meaning mentioned earlier, the time of occurrence here is also used to generate a new event sequence. As shown in Figure 7.2, when *c* is inserted into the buffer, *c* will look at the occurrence time of the event which occurred right before *c* and compose an event sequence with all events that have the same occurrence time.

In the previous version of NARS, time induction is performed on the pairs that can occur by comparing the currently selected event with all the other events in the event sequence list. Compared to such tedious steps, this new processing method is more efficient.

Other Limitations for Temporal Induction

In addition to the requirement for the sequence of events when generating the compound events, there are also strict requirements for the types of events on both sides of the implication when generating the implication relations. Implication represents a causal relationship with a corresponding temporal attribute, even though temporal induction does not consider the semantic correlation between two events, the meaning of causal relationship must be taken into account. Therefore, when the implication relation is generated, any event cannot be regarded as a pure operation, because the operation belongs to the subjective and intentional behavior of the agent, and it is inappropriate to place an operation on either sides of the implication relation.

Use a simple example to show why it is not reasonable to use pure operations on either side of implication. First of all, if there is an event “ $\langle \{Wolf\} \rightarrow [Close] \rangle$.” means there is a wolf is getting close to the agent, and an operation “ $(op(run), \{SELF\})$ ” means the agent runs away. According to the rules of time induction, the corresponding implication is

$$\langle\langle\{wolf\} \rightarrow [close]\rangle \Rightarrow (\hat{run}, \{SELF\})\rangle.$$

the idea is that if a Wolf approaches, then the agent runs away. This plausible causation does not actually have a sensible meaning. First, according to this relationship, the agent mechanically takes flight when it senses that a Wolf is approaching. Is that appropriate for the agent to run away every time when wolves are approaching? Is it reasonable for the agent to still mechanically run away if the current situation is not suitable for it? This expression is only valid if the system knows the test environment before the test and knows escape is the only option. This is not appropriate in the changing reality.

Another improper expression uses an operation as a precondition for the whole implication. For example, if there is an operation “ $(\hat{run}, \{SELF\})$ ”, and an event

“ $\langle \{SELF\} \rightarrow [Safe] \rangle .$ ” means the agent is safe, the corresponding implication is

$$\langle (\sim \text{run}, \{SELF\}) \Rightarrow \langle \{SELF\} \rightarrow [Safe] \rangle \rangle .$$

means, if the agent run away, then the agent becomes safe. This expression itself does not have semantic rationality. First of all, it does not explain why the agent should run. Second, in what circumstances to run will the agent become safe? Does running make the agent become safe under any circumstance? If the agent triggers a run operation just to get something, does that trigger this “safe” effect as well? Obviously, these questions prove the irrationality of this statement in general AI systems. Therefore, implication generation is only for two non-compound events, or to form a cognitive schema. A cognitive schema structure has three main aspects: context, action and result. The point is that if a certain context is satisfied, then taking the appropriate action will bring about a certain result. (*Drescher, 1986*). A simple cognitive schema can be expressed in Narsese as:

$$\langle (\&/, a, b) \Rightarrow c \rangle .$$

means when pre-condition “a” is satisfied, and if the system takes operation “b”, then event “c” will happen. This prevents the system from mechanizing its actions and gives meaning to each action. Because each action is taken to make a certain outcome happen, cognitive schemas are used to provide guidance for achieving goals in NARS, as described by Hammer in (*Hammer and Lofthouse, 2018*). This design allows NARS to consider the real situation every time it wants to achieve a goal, rather than mechanically selecting actions from its knowledge base and ignoring the current situation.

7.2 Anticipation

Anticipation is a process of formulating and communicating short-term expectations to sensory or motor areas (*Bubić et al., 2010*). Anticipation helps the artificial

intelligence system to predict what will happen in the near future, and prepare for the future in advance. Anticipation is also a necessary factor for triggering emotions related to future events, such as anxiety or fear. Therefore, anticipation is a very significant function for AGI systems in rapid reaction, planning, etc.

In NARS, anticipation is based on implications learned by the system, for example, if there is an implication:

$$\langle\langle\{\text{wolf}\} \rightarrow [\text{close}] \rangle \Rightarrow \langle\{\text{SELF}\} \rightarrow [\text{hurt}]\rangle\rangle.$$

the meaning of this implication is “if a wolf is getting close then NARS will get hurt”. As long as NARS learns this implication, the system should anticipate that NARS will be hurt when it detects that a wolf is approaching.

The new NARS structure extends the processing location and function of anticipation. First of all, in the old architecture, only the task selected from the buffer can generate the anticipation, so anticipation is only processed in the reasoning within main memory. This is obviously not enough, and many psychologists have also found that anticipation not only happens in a single brain region, but appears in multiple brain regions (*Bubić et al.*, 2010; *Lee and Baldassano*, 2021).

This is actually quite easy to explain, as was the case with emotions in the previous chapter. Emotions can be triggered not only by external events, but also by internal thoughts or even fantasies. The same is true for anticipation. External real events can trigger anticipation, and internal thoughts can also trigger anticipation. For example, when a person wants to do something, the person will anticipate the result, and the expected result helps the system to make a judgment about whether to take action.

Another problem is that if the anticipation based on one task has to be processed after the task is selected from the buffer, NARS may miss the occurrence of the result even if it has already happened. This is because when a large number of events come in through the sensory channels, there is no guarantee that an event will enter the buffer and be selected into the main memory in a short time and generate

anticipation. If the time interval between the premise and the result is very short in the causal relationship, then there is a certain possibility that the result has occurred, but the premise task remains in the buffer and not selected by the system. However, if the result has already happened, when the premise task is selected into the main memory and the anticipation is generated, negative evidence will be generated for the corresponding implication between the premise and the result because the result cannot be perceived.

Therefore, for external events, the new architecture moves forward the anticipation processing to the time when the event is perceived by the system. That is, when an event is successfully inserted into the buffer, and if there is an implication relationship which takes the inserted event as precondition, the post-condition will be anticipated by the system.

Such processing ensures that anticipations can be generated as soon as the precondition occurs, no matter how short the time interval between two events is, and that as long as the outcome is perceived by the system, no negative evidence will be generated against implication due to the late anticipation.

Next, this paper begins to discuss the emotion model of NARS, but before introducing the emotion model, it will first introduce the example used for testing in this project; the emotion model will then be introduced based on the example.

7.3 Aircraft Combat Game as Testing Case

The test example used in this study is an aircraft combat game. The game was developed by Boyang Xu, and is mainly used to test the ability of event handling in OpenNARS. This chapter will use the aircraft combat game to detail the changes in the new OpenNARS architecture and the reasons for the updates. It should be noted that the update and development of the new architecture and emotion modules are not only for this game. The game is just a testing case, and the development of the

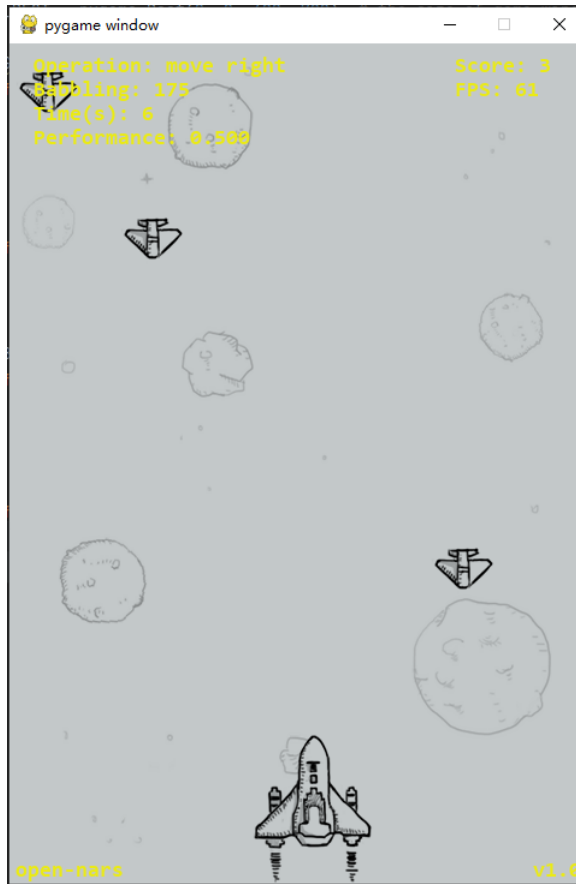


Figure 7.3: The aircraft combat game

new architecture still follows the basic assumptions of OpenNARS as an AGI system.

As shown in Figure 7.3, the game is divided into the agent’s aircraft and enemy aircraft. The goal of the game is to shoot down as many enemy planes as possible and keep them from safely passing through the bottom of the screen. The goal of the test was to let OpenNARS learn to play the game on its own and perform as well as possible.

There are four possible actions for the agent’s aircraft: move left, move right, stop, and fire. Enemy aircraft can appear from any position at the top of the screen. The agent’s aircraft perception of enemy aircraft is based on their horizontal position relative to our aircraft (left, right, or front).

After the game starts, our aircraft will perceive the position of the enemy aircraft

appearing on the screen every 250ms, convert the position information into Narsese, and send it into OpenNARS as input for reasoning; the reasoning result is fed back into the game. If the inference result is one of the above four operations, the agent's aircraft will conduct corresponding operations according to the output.

At the start of each game, there will be 200 babbling steps, during which the system will give random actions to instruct the aircraft to take different actions. By random, it means that the system does not necessarily tell the aircraft to move to the left just because there is an enemy aircraft on the left, nor does it tell the aircraft to stop and fire just because there is an enemy aircraft directly in front of our aircraft. All instructions given are initially random, and NARS will learn from all the situations that occur.

Learning process

In the learning process, all the descriptions of the environment and the behavior of the aircraft given by the system are initially random, that is to say, there is no direct relationship between NARS' perception of the environment and NARS' behavior. The primary task of NARS is to find and establish relationships among these seemingly unrelated events, although the system may be wrong.

All the relationships will be carried out in accordance with the previous section on temporal induction. There are a lot of relationships formed in the process of temporal induction. This section only introduces the induction results related to the subsequent contents.

According to the previous introduction to temporal induction, if NARS wants to trigger behavior, it depends on the relationship of a cognitive schema, that is, a context, an operation and a result. It is explained that if the agent takes the operation when the context is satisfied, it will lead to the certain result. If cognitive schematics were used to represent a possible relationship in aircraft combat games, it *might* be:

$\langle (&/, \langle \{enemy\} \rightarrow [left] \rangle, (\wedge right, \{SELF\})) \neq / \rangle \langle \{SELF\} \rightarrow [good] \rangle \rangle$.

This implication describes that if the enemy aircraft is on the left and the NARS moves to the right, then NARS could destroy the enemy aircraft and achieve its goal of being good. Here, the author deliberately uses a relation that does not follow the normal logic, since normally, when the enemy is on the left, NARS should control the plane to fly to the *left*, but here the author just wants to express that this kind of illogical knowledge is certainly possible to be generated during the process of learning. Sometimes, expectation of wrong relationship is even higher than the right one by the end of the learning process. How NARS adjusts correct and incorrect knowledge will be described in a later section.

All possible combinations will be learned in the learning process (only in this case, since the environment is simple, even so this is not guaranteed in complex environment). NARS cannot fully understand the optimal moving patterns only through 200 steps of learning, but can adjust its knowledge structure in the practice stage after the learning process. The trigger of the behavior depends on the pursuit of the goal by NARS.

In the testing phase, two methods were used. One was the game system updates the enemy aircraft position and also gives goal reminder input to NARS after the training period, and the other was the game system only updates the enemy aircraft position after the training period without the goal reminder.

7.4 Emotion Mechanism in NARS

In this section, we will look at the emotion model in NARS in detail. The introduction of this chapter will be based on the testing case of aircraft combat, and use the test case to explain the triggering of emotion and the function of emotion. The previous chapter introduced several existing emotions of NARS. Satisfaction had

already appeared in the NARS system before this study, and related explanations can be found in (*Wang et al.*, 2018), so we will not go into too much explanation of Satisfaction here.

Hope

If we review the appraisal model of hope and disappointment in the previous chapter, we will see that disappointment comes from hope, and when hope fails, disappointment will follow. First, the appraisal model of hope is for a predictable and desirable result. Bruininks describes the role of hope as keeping an eye on future outcomes and being able to foresee the approach to reach them (*Bruininks and Malle*, 2005). Similarly, in another study (*Mcdermott et al.*, 2017) describes how hope might promote a desire to seek help. Mcdermott found that people with high levels of hope were more likely to seek help than those with low levels of hope. Both studies show that hope itself causes the agent to seek ways to accomplish the goal by keeping the agent focused on the goal.

Therefore, hope diverts the agent's attention to focus on what it wants to happen. Hope is different from desire. Hope is due to the anticipation of what agent wants to happen, and it has a high degree of anticipation. However, desire is just the goal of agency, not involving anticipation.

From the above analysis, it can be seen that the desired function is to improve the priority of the relevant goal, so that the agent can transfer its attention to the relevant goal, and make corresponding actions to promote the completion of the goal according to the current situation.

There are several factors that affect the degree of hope. The initial value is determined by the desire for the goal. The more the goal is pursued, the higher the degree to which the goal is expected to be accomplished. The second is the degree of anticipation. The higher the degree of anticipation, the more likely the agent thinks

the goal will be accomplished. The final consideration is time, and the closer in time an event is to the desired goal, the higher the degree of hope. After hope is triggered, the budget value of the goal concept is calculated by the following formula:

$$P_{new} = incPriority(P_{old}, E_G, A_G, T)$$

$$D_{new} = incDurability(D_{old}, E_G, A_G, T)$$

where P_{old} and D_{old} is the original priority and durability of the goal, and E_G is the expectation of the desire value of the goal, A_G is the anticipation level, and T is the time difference between the current time and anticipated occurrence time. The inspiration for these appraisal variables mainly comes from Scherer’s work in (*Scherer et al.*, 2001). For example, E_G is like for goal relevance, and novelty is similar to A_G , and on this basis, NARS also takes time into account.

The *incPriority()* and *incDurability()* functions are internal functions for calculating the budget value, where each of them takes the “or” operation whose output is disjunctively determined by the input. For example:

$$incPriority(P_{old}, E_G, A_G, T) = (1 - P_{old}) * (1 - E_G) * (1 - A_G) * (1 - T)$$

In the testing example, the aircraft has only one goal which is “< {SELF} → [good] > .”, and the way to do that is to take out enemy planes. Suppose the NARS had been trained to move correctly, for example, if our aircraft moved to the left when the enemy aircraft was on the left, NARS could kill the enemy aircraft. The process can be expressed in Narsese as:

$$\langle (&/, \langle \{enemy\} \rightarrow [left] \rangle, (\wedge left, \{SELF\})) \Rightarrow \langle \{SELF\} \rightarrow [good] \rangle.$$

So in this situation, if the aircraft is on the left and our aircraft starts to move to the left, then the priority of the goal starts to rise, forcing our aircraft to take action in order to achieve the goal before the goal is completed.

The main function of hope is to motivate the agent to take initiative or to seek help in accomplishing the goal. Therefore, “hope” helps NARS focus on the goal it wants to achieve. The testing process used the second process mentioned earlier, in which the game system only provided position information of the enemy aircraft after the training and did not alert NARS to the goal. The purpose of such a test is to test the desired increase in NARS autonomy by prioritizing the goal concept so that NARS can focus on the goal concept. Through the selection of the goal concept in the main memory, combined with the current position of the enemy aircraft, the corresponding behavior is initiated.

Disappointment

Unlike hope, disappointment can be used to correct knowledge structure. Disappointment is the opposite of hope, disappointment happens if what is hoped by the system doesn’t happen, the greater the disparity, the greater the disappointment (*Bell, 1985*). For example, if, in the process of training, the system unfortunately learns some false knowledge, and the expectation of false knowledge is higher than that of correct knowledge, the system will preferentially select the wrong behavior when the premise event occurs. For example, during the learning process, the enemy plane is on the left, but the plane flies to the right, and destroys the enemy plane. We found that this situation is possible during the testing. Moreover, if the expectation of such knowledge is higher than the correct knowledge, the aircraft will temporarily perform the wrong behavior. However, such behavior is possible to be corrected.

In the experiment, it was found that there are several incorrect pieces of knowledge that can be possibly learned by the system. Namely, NARS flies to the right when the enemy plane is on the left, NARS flies to the left when the enemy plane is on the right, or NARS stops when the enemy plane is on the left or the right. When this happens, after the training process, NARS may get stuck on the left, because the

wrong knowledge is that NARS goes left when the enemy plane is on the right, and if the wrong knowledge is that NARS goes right when the enemy plane is on the left, then NARS may get stuck on the right. This could have happened because multiple enemy planes were on the right side of the NARS plane, but as the plane moved to the right, the bullet hit the plane that was on the left side of the plane, resulting in an erroneous observation message.

If such incorrect information is observed more often than correct information, the expectation of incorrect knowledge will be higher than correct knowledge at the end of the training process. For example, if the enemy aircraft is on the right during training, the expectation of shifting to the left is higher than the expectation of shifting to the right. If this is the case, after the training process, NARS may lead the aircraft to go all the way to the left, even to the far left of the screen, because in the knowledge base of NARS, the best option is to go left when the enemy aircraft appears on the right. However, this situation will not last long, because even if the aircraft is stuck on the left, the aircraft will detect a large number of aircraft on the right, and the NARS directive is to fly to the left, so according to the temporal induction process mentioned at the beginning of this chapter, the system will generate an event compound event:

$$(&/, \langle \{ \text{enemy} \} \rightarrow [\text{right}] \rangle, (\sim \text{left}, \{ \text{SELF} \})).$$

In the knowledge base of NARS, if the enemy aircraft is on the right, and NARS moves to the left, it will eventually make “ $\langle \{ \text{SELF} \} \rightarrow [\text{good}] \rangle$ ” to happen. Therefore, when the compound event “the enemy aircraft is on the right and the aircraft moves to the left” enters the buffer, the buffer will generate the expectation of “ $\langle \{ \text{SELF} \} \rightarrow [\text{good}] \rangle$ ”, and anticipate the result to happen at certain time.

Obviously, the chance of what is expected by the system to happen would be lower because most of the enemy aircraft would be right of our aircraft, so moving left or getting stuck at the left side won't allow the NARS aircraft to destroy the enemy.

1	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>2</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>3</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>4</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr></td></tr></td></tr></td></tr>	2	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>3</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>4</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr></td></tr></td></tr>	3	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>4</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr></td></tr>	4	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr>	5	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr>	6	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>
2	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>3</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>4</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr></td></tr></td></tr>	3	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>4</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr></td></tr>	4	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr>	5	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr>	6	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>		
3	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>4</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr></td></tr>	4	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr>	5	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr>	6	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>				
4	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>5</td> <td><(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr> </td></tr>	5	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr>	6	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>						
5	<(&/,<{enemy} --> [right]>,<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,</td></tr> <tr> <td>6</td> <td><(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>></td> </tr>	6	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>								
6	<(&/,<{enemy} --> [left]>,<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>										

Figure 7.4: Rank of Learned Knowledge at Early State of the Test Example

Therefore, if the expected event doesn't happen at expected time point, the system will generate a corresponding negative evidence for:

$$\langle (&/, \langle \{enemy\} \rightarrow [right] \rangle, (\wedge left, \{SELF\})) \rangle = / \rangle \langle \{SELF\} \rightarrow [good] \rangle.$$

with the negated truth value $(1 - f, c)$ of the anticipation, since “the greater the disparity, the greater the disappointment.”

Disappointment plays a very important role in any type of test. Because the instructions given by the system in the learning process are random, in most cases, NARS does not know the optimal way to move after the learning process. The Figure 7.4 shows the learning results of a test in the early stage.

These implications are the cognitive schemas that can achieve the goal summarized by NARS through temporal induction in the learning process, and are sorted from top to bottom by the expectation of each piece of knowledge. It is clear that the first knowledge is correct, meaning that if NARS leads the aircraft to fly to the right when an enemy aircraft is detected on the right, the goal will be achieved. However, if we look at the second implication, we can see that in the current state, if an enemy plane is detected on the left, NARS will prefer to stop rather than move to the left, because the expectation of moving to the left is lower than stop.

In the practice process, NARS will gradually reduce the expectation of the wrong knowledge by using the emotion of “disappointment” as described before. This makes the expectation value of correct knowledge exceed the expected value of incorrect knowledge. When the expected value of correct knowledge again exceeds the expected value of incorrect knowledge, the stopped NARS aircraft will start moving again because it has found a more appropriate solution. And when it starts moving

```

1 <(&/,<{enemy} --> [left],<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,
2 <(&/,<{enemy} --> [right],<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>,
3 <(&/,<{enemy} --> [left],<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,
4 <(&/,<{enemy} --> [right],<(*,{SELF}) --> ^left>) =/> <{SELF} --> [good]>>,
5 <(&/,<{enemy} --> [right],<(*,{SELF}) --> ^deactivate>) =/> <{SELF} --> [good]>>,
6 <(&/,<{enemy} --> [left],<(*,{SELF}) --> ^right>) =/> <{SELF} --> [good]>>

```

Figure 7.5: Rank of Learned Knowledge at Late State of The Test Example

again, the correct knowledge guides the aircraft with the proper response, raising the expectation of the correct knowledge. NARS adjusts the structure of its knowledge over time with the help of “Disappointment” to make right conclusions; eventually, NARS will learn the proper guidance for the game in different situations.

Figure 7.5 shows the ranking of knowledge at late state in the same test. It is clear that after a period of correction, if the enemy plane is on the left, the expectation of the “fly left” item has exceeded that of the “stop” item.

Fear

According to the appraisal model of fear, both fear and hope are triggered by future events. The difference between them is that in the appraisal model of hope, anticipation and desire have the same attitude towards the event; that is, what the agent anticipates is what the agent wants to happen. However, fear is different, as the emotional valence of fear is negative. It is precisely because of the conflict between what the agent expects and what the agent wants to happen that the goals of the agent are challenged.

So fear is handled in a similar way to hope, and it functions in a similar way, but the difference is that the two emotions are used in different situations. Hope is dealing with a positive situation, while fear is dealing with a negative situation, even though fear is meant to have a positive effect. Hope is anticipating a good outcome and trying to find a way to make it happen. Fear, on the other hand, anticipates a bad outcome and tries to find a way to prevent it.

Thus, when one event is not congruent with a goal in NARS, for example, NARS

desires an event to happen (the expectation of the desire value of the event is greater than 0.5), but NARS anticipates this event will not to happen (the expectation value of the anticipation is less than 0.45). Otherwise, if NARS doesn't desire an event to happen (the expectation value of the desire value is less than 0.5), but NARS anticipates this event will happen (the expectation value of the anticipation is greater than 0.55), fear would be triggered. It consists with the appraisal model in Lazarus theory which fear would be to be triggered whenever anticipations conflict with goals (*Lazarus, 1991*).

Fear is similar to hope in that it helps the system to shift and focus its attention on emotion-related goals. Therefore, the adjustment of the budget value of the goal concept is also similar. Compared with "hope", "fear" focuses on conflict between expectation and desire, so the greater the conflict between expectation and desire, the higher the degree of fear, and the higher the degree of adjustment of attention. Therefore, in NARS, when fear is triggered, the budget value of the target concept changes as follows:

$$P_{new} = or(P_{old}, E_G, abs(A - T)_G, T)$$

$$D_{new} = or(D_{old}, E_G, abs(A - T)_G, T)$$

Similarly with "hope", fear also takes the original budget of the concept as the starting point. E_G is the expectation of the desire value of the goal, T is the time difference between the current time and the anticipated occurrence time, and the last value $abs(A - T)_G$ is the difference between the desire value and the truth value of the anticipation.

In the testing process, "fear" sometimes goes hand in hand with "hope", because "fear" is triggered when the agent's goal is threatened in the future, so the agent is motivated to take action to prevent the goal from being threatened. However "hope" can be triggered after the agent takes action motivated by "fear", and the agent anticipates the goal will be met in some way. Therefore, the agent will continually

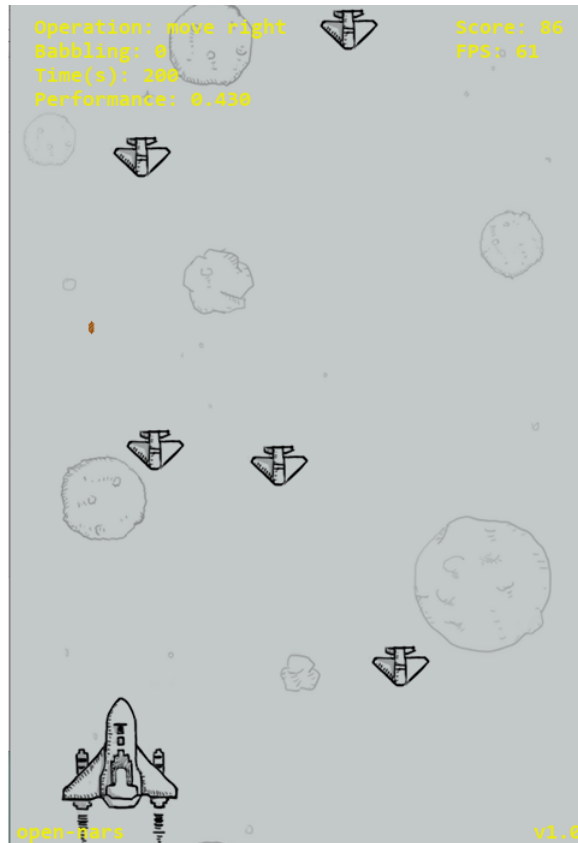


Figure 7.6: Testing Score Without Emotion After 200 seconds

focus its attention on the goal, and take action to achieve the goal to a greater extent. In other words, “fear” is triggered by having to take action to prevent an event that threatens a goal, while “hope” can be triggered by the agent seeing the possibility of achieving the goal after the agent has acted motivated by the fear.

But with the help of “hope” and “fear”, when the system detects the location of the enemy plane and knows that it can achieve the goal state through behavior, ”hope” will use this to elevate the priority and durability of the goal concept and increase the chance that the goal concept will be selected in the main memory. In this case, after each training session, the plane does not stop completely, but behaves accordingly.

In the testing with “hope” and “fear”, it is found that the priority of a goal concept can always be maintained at a higher level compared to tests without emotion. The



Figure 7.7: Testing Score With Emotion After 200 seconds

emotions increase the possibility of the goal concept being selected and processed by NARS from the concept “bag” after training. It turns out that on a 200 seconds test, the goal concept was selected at least twice as often with emotional state as with no emotional state. This is why the performance with emotion was better than the performance without emotion, since the system spent more time working to achieve the goal.

It was found that, in most cases, with the help of emotion, the arousal degree of the goal concept was higher than that of the non-emotional state. This means that the goal concept was selected more frequently in the emotional state than in the non-emotional state. Even if there is no external input to remind NARS to complete the goal, NARS can completely rely on emotion to remind itself. In most cases, the concept of the goal with no emotional state was not aroused enough to cause

the NARS to autonomously instruct the aircraft to take action after the trainer was finished without the help of external reminders. If the aircraft can't take action, then it can't correct the wrong knowledge learned in the training stage quickly. As a result, the plane will end up stuck in one place for a long time. Such a finding proves to some extent that emotion plays a significant role in the improvement of agent autonomy.

Anxiety, Relief, and Sadness

In addition to the three emotions discussed previously, the existing NARS architecture is also fully capable of achieving the remaining three emotions: relief, sadness, and anxiety. These three emotions have been implemented in the new architecture and have a certain chance to be triggered in the aircraft combat game, but due to the low complexity of the game, the three emotions in this section did not play a very significant role in the test compared to the three emotions introduced earlier. But this does not mean that these three emotions have no effect on NARS or AGI, because just like "hope" and "fear", all emotions will be tested and adjusted in a more complex environment in the future, just like the role of emotion in human life.

Anxiety and fear, like hope, are emotions triggered by anticipation of future events. In terms of valence state, anxiety and fear are the same, belonging to the negative emotion. It's easy to confuse anxiety and fear because of their negative valence and think of them as the same emotion, but in fact, there are significant differences between the two emotions. Fear is due to the explicit anticipation of an event that conflicts with a goal, whereas anxiety is an inability to make a definite anticipation of an event that conflicts with a goal (*Steimer, 2002*).

Compared with fear, expectation value of anticipation is no longer the factor that determines the degree of anxiety because anxiety means the system cannot make a confirmed expectation for the future situation. However, expectation of goal's desire value and temporal factor are still important factors that determine the level

of anxiety. More importantly, anxiety is highly possible to be triggered by an inability to determine appropriate adaptive behavior (*Selye*, 1984). This is clearly different from “fear” and “hope”, so anxiety is triggered when the agent is unable to accurately anticipate the outcome and does not have the appropriate behavior to deal with the possible outcome. For NARS, the role of anxiety is similar to “fear” and “hope”, as it is a way of focusing attention to goal concepts which cause the agent to feel anxiety. Because of the particularity of anxiety, anxiety may not promote any behaviors, since there is no confirmed solution to solve the problem, but it can keep the priority of a goal concept at a certain high level, which prompts the agent to ask questions or generate sub-goals to find a final solution to the problem.

Anxiety in the human brain involves more triggering conditions and factors, while anxiety in NARS only considers how anxiety can help the system from the perspective of function, without considering a complete model simulating human anxiety. Anxiety, of course, didn’t play a significant role in the test, because there were no moments in the game when NARS didn’t know what to do, even though NARS was instructing the plane to do the wrong thing, and there was no feeling overwhelmed. However, as testing environments become more complex, anxiety will play a role in future testing.

In contrast to all emotions mentioned, relief and sadness are not triggered by something that will happen in the future, but by something that has already happened. Relief can correspond to a variety of emotions, such as fear and anxiety. Relief is triggered when the event anticipated by fear or anxiety is confirmed to have not happened.

In the new architecture, the buffer has a special module that removes expired expectations. Disappointment also occurs in this module. When the anticipated event does not occur at the expected time point, the expectation will be deleted, and a negative evidence for the corresponding implication will be generated against the

original implication through disappointment, which will be used to update the belief of NARS through revision in the main memory. Disappointment corresponds to an expected event that did not occur, but relief is generated if an event opposite to the expected event occurs and the event occurs in accordance with the goal.

The main effect of relief is to release the resources that have been focused by fear or anxiety. Since the crisis has been relieved, there is no longer a need to focus attention or cognitive resources on the cause of the fear or anxiety. Since relief is used to release resources, in NARS it lowers the priority of the corresponding target concept. The formula is as follows:

$$P_{new} = decPriority(P_{old}, abs(A_G - T_G), T)$$

$$D_{new} = decPriority(D_{old}, abs(A_G - T_G), T)$$

where P_{old} and D_{old} is the original priority and durability, which is the budget value raised by fear or anxiety, T_G is the truth value of the confirmed event, and A_G is the truth value of the anticipation. Therefore, the bigger the difference between the confirmed event and anticipated event, the more attention will be relieved. The last factor T is still the time, as the degree of relief will get higher and higher over time.

Here the $decPriority()$ function is also a internal budget value function which takes an “Operation”. The output of the “and” operation is conjunctively determined by the inputs. For example:

$$decPriority(P_{old}, abs(A_G - T_G), T) = P_{old} * abs(A_G - T_G) * T$$

The primary function of relief is to free up resources by prioritizing concepts so that the system’s attention is not focused on concepts that are activated by fear or anxiety. Relief was not included in the aircraft combat game because the aircraft was in a high wake-up state from the start to the end of the game, and there was no respite for our aircraft as the enemy planes flew in from above the screen.

The final emotion in this article is sadness. Sadness is actually opposite to satisfaction. First of all, both are aimed at confirming what has happened, rather than what to happen in the future. Satisfaction means that the reality is getting closer and closer to the desired situation, while sadness means that the events that have happened are far from the ideal state.

In addition to NARS trying to be satisfied, NARS also avoids situations that make it feel sad. This avoidance behavior also coincides with fear, because fear motivates the individual to take action to avoid an expected event which the agent does not want to happen. This also explains what Zhan pointed out in (*Zhan et al.*, 2017), that sadness has a promoting effect on fear.

Although the three emotions described in this section did not make a significant contribution to the testing procedures used in this article, this does not mean that these three emotions are useless for NARS. Of course, this issue is worth further exploration. In future works, we should not only explore under what environment emotion can help NARS improve adaptation and reasoning ability, but also explore what role emotion plays in a more complex environment.

7.5 Summary

This chapter introduced the main implementation work of the whole doctoral research. It includes the realization of new architecture, the realization of emotion module, and the related experiments.

Figure 7.8 shows the internal view of the buffer, describing both the Overall Experience buffer and the Internal Experience buffer. The difference is that for the Overall Experience buffer, the input can be either external input or inference from the Internal Experience buffer. However, only external input can trigger emotion in Overall Experience buffer, because the Internal Experience buffer has already emotionally processed the information from Memory.

When the external input enters into the buffer, the information first enter the priority list as described earlier, along with the event sequence. When the event successfully enters the event sequence (it was proved that the the event was successfully inserted into the priority list), it is shown that the event is perceived by NARS as a confirmed event. Therefore, the event will be evaluated through the emotion module of sadness and relief. In addition to evaluating events, events also generate expectations from anticipations based on what has been learned and places expectations into an anticipation list. The expectation list stores what NARS expects to happen in the future. Therefore, hope, fear and anxiety are evaluated by their respective appraisal after the anticipation is generated and corresponding emotions are generated.

Each working cycle requires the highest-priority task to be pulled from the priority list, and each time it does, the system traverses the anticipation list and removes expired anticipation that did not occur at the expected point in time. In the process of deletion, the system will be disappointed because the anticipation fails, and negative evidence will be generated against the implication of the expectation to lower the truth value of the corresponding causality.

All of these emotional functions will be aggregated in memory and will have an impact on the cognitive resources in conceptual memory. Through aircraft combat tests, it was shown that emotion does indeed play a significant role in the application of NARS.

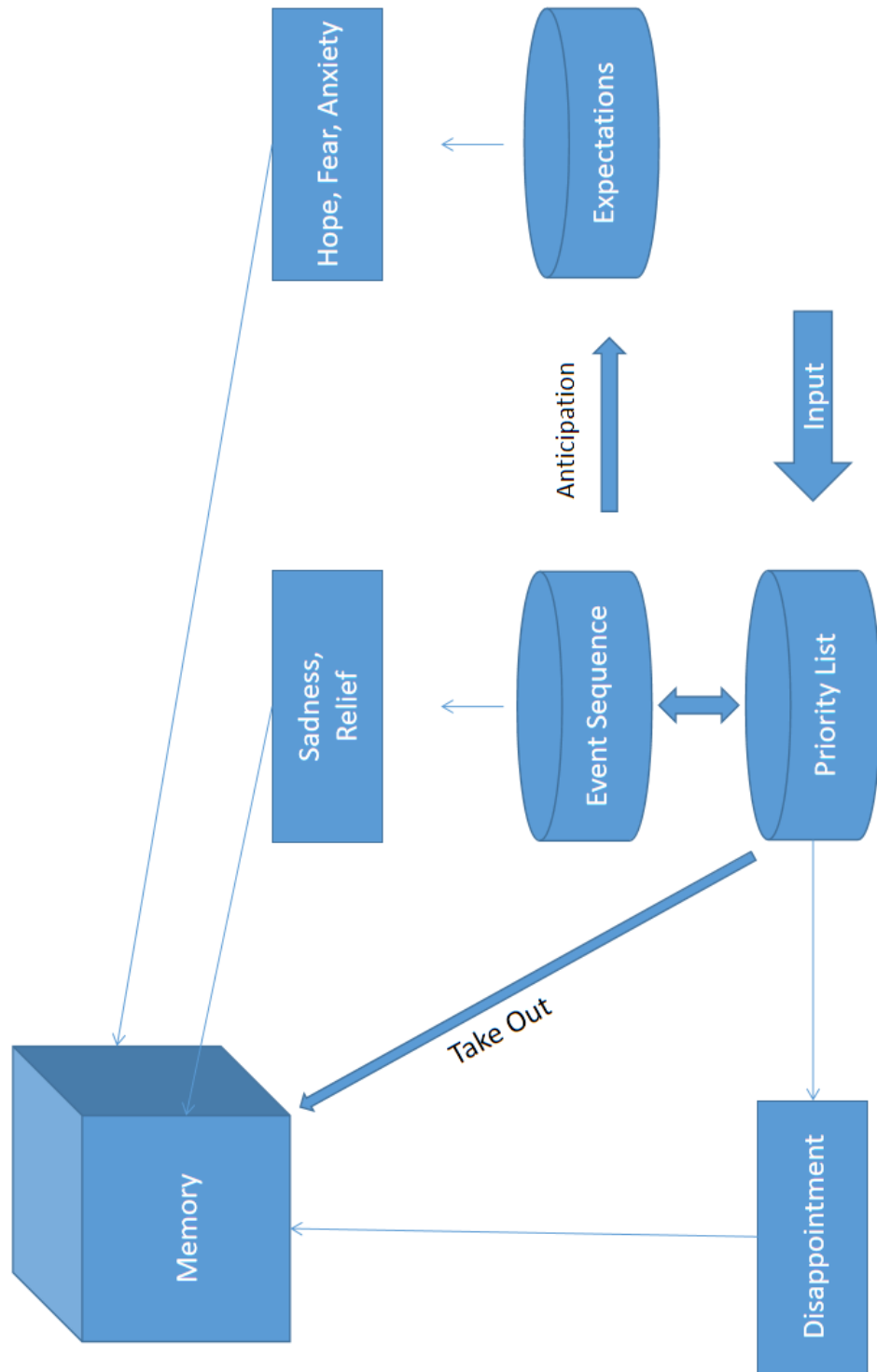


Figure 7.8: Internal view of the Buffer

CHAPTER 8

COMPARISON WITH EXISTING ARCHITECTURES

This chapter will review and compare the work on emotion introduced in Chapter 3. It is important to note that the purpose of this chapter is to compare existing works on emotion with the work in NARS. There is no intention to judge if a work is good or bad.

8.1 Affective Computing

Affective computing emphasizes the recognition function of machines to human emotion in human-computer interaction. This includes recognition of facial expressions, postures, movements, language and tone, judging the emotional state of the other side through these external manifestations, and then making correct feedback for different emotional states. The above criteria are also the most commonly used criteria for human interaction. In addition to the above judgment criteria, Affective computing is trying to make the detection with more sensors, such as sensors for temperature, blood pressure and other factors, to further judge the emotional state of the other party. Of course, these sensors can also be used as judgment criteria when the other party has no obvious external behavior.

In addition to external representations, the emotion understanding module of affective computing will attempt to analyze the causes of the emotional states of the interactors to further determine the emotional states of the interactors. In the past

few years, deep learning has provided a lot of support for the emotion recognition function of emotion computing. Different researchers have provided outstanding emotion recognition methods from different perspectives (*Santhoshkumar and Geetha*, 2019; *Khan et al.*, 2021; *Hossain and Muhammad*, 2018). And these different methods constitute the big frame of affective computing.

Therefore, affective computing is a multi-dimensional computer model of emotion processing. The main function is human-computer interaction. In general, the functions and basic assumptions of affective computing are quite different from those of NARS. Human-computer interaction is not regarded as the primary function to be solved in NARS, the emotion in NARS is mainly used for the adjustment of internal resources and knowledge structure of NARS, and does not involve the judgment of other people's or intelligent system's emotional state.

Human-computer interaction isn't just a phenomenon in deep learning or mainstream AI research. There are also similar studies in the field of general artificial intelligence. In (*Barone et al.*, 2008), Barone introduced an emotion model aiming at human-computer interaction. The task of the robot is to guide tourists, and the robot will generate emotions according to external stimuli, adjust internal resources, and generate corresponding behaviors to complete its goals. Therefore, affective computing and emotion in AGI are not separate research fields. Although affective computing deals with emotions differently than NARS, but it is undeniable that the research on the human-computer interaction function of affective computing provides inspiration and help for the future research on the human-computer interaction of AGI.

8.2 ACT-R

In the study of emotion using ACT-R, it is found that the mechanism provided in the framework of ACT-R can be used to simulate several different emotion models. This is in line with the requirements mentioned in Chapter 5. First, emotion modules

are part of the framework, not external plug-ins. Second, emotion modules should be implemented by utilizing the functions of the cognitive framework itself. The implementation method of the emotion module in ACT-R verifies the rationality of the design requirements and architectural requirements in Chapter 5.

In terms of implementation methods, NARS and ACT-R are similar in that both frameworks take into account the importance of the goal, valence, certainty and other factors. However, ACT-R is more likely to use numerical combination of various factors to build mapping to emotions. However, the appraisal model in NARS uses the relationship between events and goal of NARS for cognitive appraisal, and the value of the corresponding relationship will affect the degree of arousal. From the perspective of function, research on emotion using ACT-R involves the influence of emotion on decision making and memory, and in this respect, the research direction of ACT-R on emotion is consistent with that of NARS.

In general, the research objectives of ACT-R are different from those of NARS. Compared with affective computing, ACT-R does not take human-computer interaction as its main research direction, but it is designed to simulate human cognitive theory. The research purpose of NARS is not to simulate any specific human cognitive theory, but to improve the adaptive ability of NARS and the various functions of NARS under the guidance of different theories of human intelligence. The research of NARS is not to make functions of NARS or the behaviors of NARS more like human beings, but to combine the respective advantages of human and machine, and to create a cognitive framework with a unique form of intelligence inspired by human intelligence.

8.3 CLARION

Although CLARION is very different from NARS in the implementation of the framework, CLARION shares a lot of similarities with Nars in the emotional process-

ing. CLARION is a cognitive framework composed of different sub-modules, but in NARS all functions are intertwined, and a single function is represented in the form of an independent module. Although I always use word “emotion module” in the previous discussion, however, as can be seen from the internal structure of the buffer, there is no single module for special processing of emotion. Different appraisal and processing models are placed in different parts according to the appraisal requirements. When I mention the word ”emotion module”, it means the synthesis of all parts that can deal with emotion.

CLARION also regards emotion as the result of the operation of the whole cognitive system, and the triggering of emotion is the mutual result between internal needs and external environmental factors. In NARS, this statement can be understood as the relationship between the goal of NARS and the external stimulus. CLARION, like NARS, regards emotion as the result of “cognitive appraisal”.

CLARION also has a continuous cycle of appraisal and re-appraisal , meaning that a cognitive and behavioral response to an emotion can serve as an input for further appraisal. This functionality is also discussed in Chapter 7, when we discussed fear and hope. In testing, the implication which has a goal as the post-condition, the premise is most likely to be a composite event consisting of the context and the action. The meaning lies in that only when the context is satisfied, the goal can be satisfied if the action is taken. For example, when an enemy plane appears on the right side of the plane, the goal can be achieved by steering the plane to fly to the right. Therefore, when NARS takes actions due to fear to avoid the threat to the goal, the actions taken can form a compound event with the context, and then form the expectation for the goal achievement, and this anticipation satisfied the appraisal model of hope.

The only difference in perspective is that CLARION argues that emotions are closely linked to action, and emotions usually lead to action. In the emotional pro-

cessing of NARS, there is no such relationship hypothesis. The emotional processing of NARS will eventually fall on internal resource regulation, and the result of internal resource regulation is the diversion of attention to a goal concept, but this does not guarantee a behavior is going to be triggered. Emotion does offer motivation to some behavior, but most time, people will choose to receive motivation or reject motivation through rational thinking. For example, anger will prompt some destructive behavior, but many people tend to control the desire to carry out destructive behavior. Therefore, in NARS, there is no mapping relationship between emotion and behavior. Emotion only regulates internal resources or knowledge structure and does not guarantee the generation of behavior.

Nevertheless, CLARION has a lot in common with NARS, and a lot of ideas that NARS can learn from. For example, CLARION uses coping potential as one of the evaluation criteria, and coping potential is an important appraisal criterion in Lazarus's cognitive theory (*Lazarus, 1991*), but NARS is not capable of dealing with coping potential at this stage, but it will be considered in the future work

8.4 LIDA

LIDA and NARS have similar inspirations for emotion processing, such as Lazarus' emotion appraisal model (*Lazarus, 1991*) and Scherer's emotion intensity calculation model (*Scherer et al., 2001*). However, there are still great differences in many processing details.

The emotion appraisal models in NARS are placed at various locations in the program, such as those described in the previous section, where future-related processing is part of the anticipation process, and present or past related processing is part of the event perception process, and in addition to events, "thoughts" or "imagination" of NARS can trigger emotions. LIDA has a different appraisal Codelets and emphasizes the role of emotion valence, while the emotion valence of the emotion event is the

sum of the emotion valence of the emotion nodes related to the event. The positive and negative of the sum of the emotion valence determines the increase or decrease of the significance of the basic incentive. However, NARS is more concerned about the abstract relationship between the event and the goal of NARS, and the degree of motivation is determined by the degree of influence of other factors on the goal.

8.5 MicroPsi

MicroPsi is closer to human emotions than NARS in emotion processing because it is built based on a psychological theory. While many of the features in NARS were built with psychological inspiration, that doesn't mean the goal of design is to get closer to the way that happens in human cognition. MicroPsi has many built-in needs, psychological needs, social needs, and cognitive needs, and a goal in MicroPsi is defined as a situation to be satisfied. However, NARS does not have any built-in needs or goals which need satisfied; any goal is given by the outside world on real time during the interaction with the environment, or is generated by the goal as a sub-goal. Therefore, when NARS is really placed in an environment, NARS does not know what event will trigger emotions, and the emotional events that will trigger emotions are also learned in the process of interacting with the environment.

Beside differences, NARS and MicroPsi are similar in many aspects, both have appraisal models for different emotions, even though the events that trigger emotions are different. Motivational systems, in which motivation is processed as goals in NARS, but built-in needs and goals in MicroPsi. Emotion is also seen as the result of the combination of motivation and perceptual representations, and in terms of emotional function, both of them regulate internal resources through emotional arousal.

8.6 Sigma

Sigma’s design of cognitive architecture provides a lot of inspiration for this paper to propose the requirements of constructing emotion modules in AGI system. In line with the NARS design philosophy, the generality of the system cannot be lost on the basis of the continuous expansion of functionality from rigorous cognitive processing to the ability to handle complex events in the real world. Although the function of NARS’s emotions is different from that of Sigma in previous studies, the processing of emotions in this study has many similarities with Sigma.

Limited resources have always been the primary problem facing AGI systems, and this is a fundamental assumption that cannot be ignored under any circumstances. The emotion in NARS and the emotion in Sigma are used to make reasonable resource allocation in the limited resources to improve the system’s adaptability and reasoning ability.

8.7 NARS

Before this study, Wang et al. briefly introduced some works related to emotion in NARS in (*Wang et al.*, 2016), which has many similarities and differences with the work described in this paper. First of all, both work identified the need for emotion in the AGI system, and both use emotion to enhance the ability of the system to manage resources under the condition which ”knowledge and resources are relatively insufficient”, and this ability is not a complete simulation of the corresponding functions IN the human brain. two works also emphasize the necessity of emotion in solving multiple goals or even goal conflict, and the regulation of emotion on the machine plays a role in the allocation of internal resources.

The difference between the two lies in that the study in this dissertation integrates different emotions into the original framework of NARS according to different

appraisal criteria in cognitive emotion theory, and makes corresponding adjustments to the internal cognitive resources based on different emotions. However, the previous elaboration on emotion is not to directly add emotion to the system, but different functions of NARS which related to tasks, goals and internal resource adjustment reflect the relevant performance of human emotion. However, the difference between the two is a chicken-and-egg problem. The effectiveness and functionality of the two methods will be reflected in the future with the development of NARS.

8.8 Summary

This chapter briefly compares the concepts of framework design and functional design of emotion modules in NARS with the existing cognitive models. Obviously, there is a big difference between AGI and mainstream AI in the approach of pursuit emotions, due to different assumptions about generality and insufficient resources. Mainstream AI does not pursue the generality of the system, but strives to achieve better performance in a certain function. Moreover, the mainstream does not consider the effectiveness of resources, so it does not need the emotion to help the system to make appropriate resource allocation on the limited resources. Although such differences do not allow the two to make too much comparison in the realization of emotional functions, it does not mean that they cannot cooperate in the realization of emotional functions.

Different AGI systems have slightly different approaches to emotion processing, but from a functional point of view, most emotion processing focuses on the awakening and adjustment of attention, cognitive resources. In most AGI systems do not take human-computer interaction as the primary goal of emotional processing, but this is also a problem AGI systems will have to face in the future. In the future, emotional processing in affective computing on the human-computer interaction process will provide much inspiration and support for AGI's human-computer interaction, at the

same time, AGI's emotional design can also provide invaluable assistance to AI in resource allocation, decision making, and other cognitive functions.

CHAPTER 9

CONCLUSION AND FUTURE WORKS

This chapter will provide a summary of NARS' work on emotion, a statement of the known limitation on emotion module. Finally, a description of the future plans on emotion works will be introduced.

9.1 Major Results

Emotion, as a complex cognitive process, plays an important role in human life, learning, social interaction and so on. The theoretical and practical contributions of this paper first answer one question: Can AI have emotions?

Through the theoretical elaboration and experiment of this paper, it is found that the general artificial intelligence systems have the potential to possess emotion. It is important to note, however, that the emotions of AGI do not have to and will not be exactly the same as those of humans. Although the cognitive framework and emotion framework of the artificial general intelligence system are based on human psychology, cognitive science and neuroscience as the theoretical basis and inspiration sources, it does not mean that the emotion of the general artificial intelligence system needs to take human emotion as the ultimate goal. It's like the airplane is a machine inspired by the shape of a bird, but the airplane contributes more to humanity from a functional point of view than the bird does to us.

Therefore, it is not appropriate to evaluate machine emotion by human emotion, because at present, the research of AGI on emotion is to explore the effect of emotion on the thinking process, the attention shift and adaptability improvement of intel-

ligence systems, based on the basic assumption that "knowledge and resources are relatively insufficient". However, from the perspective of function, this paper proves through experiments that emotion plays an important role in the adjustment of cognitive resources, attention shift and improvement of autonomy for general artificial intelligence. Therefore, the answer to the above question is given: machines can have emotions, and emotions are necessary and important for general intelligence systems.

Based on the research on emotion in NARS, this paper puts forward the requirement of building emotion module in intelligence system, the idea comes from the conceptual design of NARS and other general artificial intelligence, as well as the common point of emotion function in different general artificial intelligence systems. These requirements are not guaranteed to be applicable to all intelligence systems, but can be applied to most intelligence systems with the same design philosophy. The research on emotion of general artificial intelligence is still in its infancy, the requirements for emotion design of general artificial intelligence in the future will not be limited to the requirements mentioned in this paper.

In this study, the new architecture of NARS is implemented, and the corresponding modifications are made according to different problems in the implementation process. On this basis of the new architecture, the emotion module is added. The emotion module of NARS is not targeted at any particular task, or pre-defined goals or needs. The appraisal model focuses on the abstract relationship between the event and the goals in NARS, so the emotion module of NARS remains general.

Although the appraisal model of emotion is innate in the new architecture of NARS, this does not affect the generality of NARS as well as the emotion modules. NARS still needs to learn what will trigger emotion through interaction with the environment.

During the test of the game, "hope", "fear" and "disappointment" were triggered as expected in the game. It's worth knowing that NARS knew nothing about the

game prior to the test. There is no special module or system that dictates what emotion NARS should trigger during the game. Therefore, in the testing process of this game, it was proved that new emotion module of NARS has the generality, as well as its functional integrity. Functional integrity does not mean that the design and function of the emotion module in NARS has been perfect, but that the research objectives of this study have been completed.

Although some emotions have been realized in NARS, not all have not been triggered in the test process. However, with the excellent performance of the above three emotions in the revision of NARS knowledge structure and improvement of autonomy, emotions that have not played a role for the time being will play a role in the more complex test environment in the future.

This study is a preliminary attempt to realize complex emotions in NARS, and the temporary success also provides confidence and points out the direction for future research in this field. However, it can not be ignored that there are some unsolved problems in the current design.

9.2 Limitations and Future Works

Even with the temporary successful implementation of complex emotions in NARS, some problems still emerged during the testing process, and these problems should not be ignored in future design.

As Belavkin mentioned in (*Belavkin, 2001*), When emotion arouses a concept in surrogate memory, an appropriate degree of arousal will help in learning and reasoning. Just as the arousal of “hope” and “fear” to the goal in the test, NARS can trigger actions to achieve the goal by relying on internal arousal alone without the help of external input. But again, a high or low arousal level can hinder learning and reasoning. In general, at the beginning of the training process, the goal has just been realized by the agent, the degree of arousal will not be very high, so the frequency of

active selection by NARS in memory will maintain in low level; However, there are times during training when arousal levels were very high, and external input being to conflict with the voluntary behavior triggered by emotional arousal, especially when NARS first learns the wrong way to act. For example, NARS first learned that when the enemy plane is on the left, it moves to the right. At this point, due to the high arousal, NARS is constantly performing the move to the right autonomously, which causes the plane to get stuck on the right, even though the external input is trying to move the plane to the left. But it will still be forced back to the right by autonomous actions of NARS.

Even though this problem has since been solved by setting an activation threshold that triggers activation when the goal concept's priority falls below a certain level (the system still triggers emotions), it's not clear that this is the right way to solve this problem. In future studies, the degree of activation will be taken into consideration to ensure that the arousal degree of each emotion to NARS is kept within a reasonable range, so as to ensure that emotion plays a helpful role in NARS, rather than a hindrance.

Another problem is that the appraisal model of the NARS emotion module is still incomplete, which, of course, depends on the current level of NARS completion. For example, in Lazarus's cognitive theory of emotions, there are three levels of appraisal, and now NARS can do a second level of appraisal, and emotions that require a third level of appraisal are not currently assessed in NARS. In comparison with general artificial intelligence systems with emotional design such as CLARION, coping potential has been mentioned as an important appraisal criterion which not implemented in the current design of NARS. There are other standards in addition to coping potential, and these standards and capabilities will be expanded as NARS continues to improve.

The tests used in this study were not very complex, which is why some emotions

didn't contribute much in the test case. And emotion in human life, often help human survival or improve the ability to adapt in the complex environment. Moreover, success in a simple environment does not completely guarantee that the current design can still help NARS in a complex living environment, as the first question in this chapter introduces. Of course, with the deepening of this research in the future, NARS will also be tested in more complex environments, and the emotional system of NARS will continue to be improved in these tests.

The abusing emotions is a major problem that emotion machines will face. How to use emotion reasonably to guide appropriate behavior is very important for both people and emotion machines. For example, one possibility that people have a fear of emotional machines comes from what an AI system should do if it gets "angry". Controlling something like anger is not about getting rid of anger, but educating machines to control "anger", and one way to do that is to use "emotional conflict". When machines are angry and want to behave in a destructive way, anticipation of the consequences of destructive behavior can trigger "fear" that inhibits destructive behavior. In addition to emotional conflict, it is important to educate machines to not behave in a destructive way even when they are "angry", or how to suppress "anger" subjectively, which will be the focus of future AGI emotion research.

Although human-computer interaction is not currently the main goal of emotion research in NARS, but it will become the main research goal in the future. A machine with capability of cognitive appraisal is fully capable of producing "empathy" with a human, because the appraisal model in NARS is able to enable NARS to understand the cause of the another's emotional state, and can "empathize" it into its own emotion appraisal model. Such "empathic" ability will motivate NARS to make correct feedback behavior. "Empathic" ability will greatly improve NARS 'communication ability in human-computer interaction in the future, and improve its adaptability in real life.

In conclusion, this study is only the beginning of the exploration in the field of AI emotion research. The results of this study also see the dawn of artificial intelligence emotion research, and confirmed the necessity of emotion in artificial intelligence system, as well as the possibility of emotion realization in artificial intelligence system. The research on artificial intelligence emotion will be continued all the time. While continuously expanding the emotional function in artificial intelligence system, it is hoped that it can provide help for the pursuit of emotion research in different disciplines in the future.

BIBLIOGRAPHY

- Al-Shawaf, L., D. Conroy-Beam, K. Asao, and D. Buss (2015), Human emotions: An evolutionary psychological perspective, *Emotion Review*, pp. 173–186.
- Antos, D., and A. Pfeffer (2011), Using emotions to enhance decision-making., in *IJCAI International Joint Conference on Artificial Intelligence*, pp. 24–30.
- Arbib, M. A., and J. M. Fellous (2004), Emotions: From brain to robot, *Trends in Cognitive Sciences*, 8(12), pp. 554–561.
- Arnold, M. B. (1950), An excitatory theory of emotion, in *Feelings and emotions; The Mooseheart Symposium*, pp. 11–13.
- Baars, B. J. (2017), *The Global Workspace Theory of Consciousness*, chap. 16, pp. 227–242, John Wiley & Sons, Ltd.
- Bach, J. (2003), The MicroPsi Agent Architecture, in *Proceedings of ICCM5 International Conference on Cognitive Modeling Bamberg Germany*, vol. 1, pp. 15–20.
- Bach, J. (2012), A framework for emergent emotions, based on motivation and cognitive modulators, *Int. J. Synth. Emot.*, 3(1), pp. 43–63.
- Bach, J., M. Coutinho, and L. Lichtinger (2019), *Extending MicroPsi’s Model of Motivation and Emotion for Conversational Agents*, pp. 32–43.
- Barone, R., I. Macaluso, L. Riano, and A. Chella (2008), A brain inspired architecture for an outdoor robot guide, in *AAAI Fall Symposium - Technical Report*, pp. 27–34.
- Barrett, L. (2017), *How Emotions are Made: The Secret Life of the Brain*, Houghton Mifflin Harcourt.
- Belavkin, R. (2001), The role of emotion in problem solving, in *Proceedings of the AISB’01 Symposium on Emotion, Cognition and Affective Computing*, pp. 49–57.
- Belavkin, R. (2003), On emotion, learning and uncertainty: A cognitive modelling approach.
- Bell, D. E. (1985), Disappointment in decision making under uncertainty, *Operations Research*, 33(1), pp. 1–27.
- Bruininks, P., and B. Malle (2005), Distinguishing hope from optimism and related affective states, *Motivation and Emotion*, 29, pp. 324–352.

- Bubić, A., D. Cramon, and R. Schubotz (2010), Prediction, cognition and the brain, *Frontiers in human neuroscience*, 4, Article 25.
- Chown, E., R. Cochran, and F. Lee (2006), Modeling emotion: Arousal’s impact on memory., *Proceedings of the Annual Meeting of the Cognitive Science Society*, 28.
- Citron, F., M. Gray, H. Critchley, B. Weekes, and E. Ferstl (2014), Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework, *Neuropsychologia*, 56, pp. 79–89.
- Cowan, N. (2008), What are the differences between long-term, short-term, and working memory?, *Progress in brain research*, 169, pp. 323–338.
- Crockenberg, S., and E. Leerkes (2012), Infant social and emotional development in family context, *The Handbook of Infant Mental Health*, pp. 60–90.
- Damasio, A. (1999), *The Feeling of what Happens: Body and Emotion in the Making of Consciousness*, Harvest book, Harcourt Brace.
- Damasio, A. (2011), *Self Comes to Mind: Constructing the Conscious Brain*, Random House.
- Darwin, C. (2013), *The Expression of the Emotions in Man and Animals*, Cambridge Library Collection - Darwin, Evolution and Genetics, Cambridge University Press.
- Deak, A. (2011), Brain and emotion: Cognitive neuroscience of emotions, *Rev Psychol*, 18, pp. 71–80.
- Denham, S., M. Salisch, von, T. Olthof, A. Kochanoff, and S. Caverly (2002), Emotional and social development in childhood, *Blackwell Handbook of Childhood Social Development*, pp. 307–328.
- Drescher, G. (1986), Genetic ai: Translating piaget into lisp, *Instructional Science*, 14, pp. 357–380.
- Ekman, P. (2009), Darwin’s contributions to our understanding of emotional expressions, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364, pp. 3449–3451.
- Elster, J. (2009), Emotional choice and rational choice, in *The Oxford Handbook of Philosophy of Emotion*, edited by P. Goldie, Oxford University Press.
- Franklin, S., T. Madl, S. D’Mello, and J. Snaider (2014), Lida: A systems-level architecture for cognition, emotion, and learning, *IEEE Transactions on Autonomous Mental Development*, 6(1), pp. 19–41.
- Fu, X. (2016), *Psychology of Emotion*, East China Normal University Press.
- Gadanho, S. C. (2003), Learning behavior-selection by emotions and cognition in a multi-goal robot task, *J. Mach. Learn. Res.*, 4, pp. 385–412.

- Gavrilov, A. (2008), Emotions and a prior knowledge representation in artificial general intelligence, Institute of Information Theories and Applications FOI ITHEA.
- Halstead, W. C. (1961), Emotion and personality. vol. 1, psychological aspects. vol. 2, neurological and psychological aspects. magda b. arnold. columbia university press, new york, 1960., *Science*, 133(3451), pp. 455–455.
- Hammer, P., and T. Lofthouse (2018), *Goal-Directed Procedure Learning: 11th International Conference, AGI 2018, Prague, Czech Republic, August 22-25, 2018, Proceedings*, pp. 77–86.
- Hammer, P., T. Lofthouse, and P. Wang (2016), The OpenNARS Implementation of the Non-Axiomatic Reasoning System, in *Artificial General Intelligence*, pp. 160–170.
- Hammond, M. (2006), *Evolutionary Theory and Emotions*, pp. 368–385, Springer US, Boston, MA.
- Hanania, R., and L. Smith (2010), Selective attention and attention switching: Toward a unified developmental approach, *Developmental science*, 13, pp. 622–35.
- Hossain, M. S., and G. Muhammad (2018), Emotion recognition using deep learning approach from audio-visual emotional big data, *Information Fusion*, 49, pp. 69–78.
- Izard, C. (1968), The emotions as a culture-common framework of motivational experiences and communicative cues.
- Izard, C. E. (1991), *The psychology of emotions*, Emotions, personality, and psychotherapy, Plenum Press, New York.
- James, W. (1884), What is An Emotion?, *Mind*, os-IX(34), pp. 188–205.
- Joseph, C., M. Donna, K. Rosanne, and C. Rosemary (1994), A functionalist perspective on the nature of emotion, *Monographs of the Society for Research in Child Development*, 59(2/3), pp. 284–303.
- Kauschke, C., D. Bahn, M. Vesker, and G. Schwarzer (2019), The role of emotional valence for the processing of facial and verbal stimuli—positivity or negativity bias?, *Frontiers in Psychology*, 10, Article 1654.
- Khan, A. N., A. A. Ihalage, Y. Ma, B. Liu, Y. Liu, and Y. Hao (2021), Deep learning framework for subject-independent emotion detection using wireless signals, *PLOS ONE*, 16(2), e0242,946.
- Kleinginna, P., and A. Kleinginna (1981), A categorized list of emotion definitions, with suggestions for a consensual definition, *Motivation and Emotion*, 5, pp. 345–379.

- LaGrandeur, K. (2015), Emotion, artificial intelligence, and ethics, in *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*, pp. 97–109, Springer International Publishing, Cham.
- Lazarus, R. (1991), *Emotion and Adaptation*, Oxford University Press.
- Lee, C., and C. Baldassano (2021), Anticipation of temporally structured events in the brain, *eLife*, 10.
- Li, X., P. Hammer, P. Wang, and H. Xie (2018), Functionalist emotion model in nars, in *Artificial General Intelligence*, pp. 119–129, Springer International Publishing, Cham.
- Maranon, G. (1985), Contribution to the study of the emotive action of adrenaline, *Studies in Psychology*, 6(21), pp. 75–89.
- Matsumoto, D., and B. Willingham (2009), Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals, *Journal of personality and social psychology*, 96, pp. 1–10.
- McCall, R., S. Franklin, U. Faghihi, J. Snaider, and S. Kugele (2020), Artificial motivation for cognitive software agents, *Journal of Artificial General Intelligence*, 11, pp. 38–69.
- Mcdermott, R., H.-L. Cheng, Y. J. Wong, N. Booth, Z. Jones, and T. Sevig (2017), Hope for help-seeking: A positive psychology perspective of psychological help-seeking intentions, *The Counseling Psychologist*, 45, 001100001769,339.
- Nesse, R. (1990), Evolutionary explanations of emotions, *Human Nature*, 1, pp. 261–289.
- Ortony, A., G. L. Clore, and A. Collins (1988), *The Cognitive Structure of Emotions*, Cambridge University Press.
- Panksepp, J. (2010), Affective neuroscience of the emotional brain mind: Evolutionary perspectives and implications for understanding depression, *Dialogues in clinical neuroscience*, 12, pp. 533–45.
- Park, S., and R. Myung (2012), A conceptual framework for emotional response of product with act-r cognitive architecture, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), pp. 1020–1024.
- Picard, R. W. (1997), *Affective Computing*, MIT Press, Cambridge, MA, USA.
- Picard, R. W. (2003), Affective computing: challenges, *International Journal of Human-Computer Studies*, 59(1), 55–64, applications of Affective Computing in Human-Computer Interaction.
- Plutchik, R. (1982), A psychoevolutionary theory of emotions, *Social Science Information*, 21(4-5), pp. 529–553.

- Roseman, I. J. (1996), Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory, *Cognition and Emotion*, 10(3), pp. 241–278.
- Roseman, I. J. (2011), Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses, *Emotion Review*, 3(4), 434–443.
- Rosenbloom, P., A. Demski, and V. Ustun (2016), The sigma cognitive architecture and system: Towards functionally elegant grand unification, *Journal of Artificial General Intelligence*, 7.
- Rosenbloom, P. S., J. Gratch, and V. Ustun (2015), Towards emotion in sigma: From appraisal to attention, in *Artificial General Intelligence*, pp. 142–151, Springer International Publishing, Cham.
- Russell, J. (1983), Pancultural aspects of the human conceptual organization of emotions, *Journal of Personality and Social Psychology*, 45, pp. 1281–1288.
- Russell, J. A. (1989), Chapter 4 - measures of emotion, in *The Measurement of Emotions*, edited by R. Plutchik and H. Kellerman, pp. 83–111, Academic Press.
- Santhoshkumar, R., and M. K. Geetha (2019), Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks, *Procedia Computer Science*, 152, pp. 158–165.
- Sauter, D., O. Crasborn, T. Engels, R. Kamiloglu, R. Sun, F. Eisner, and D. Haun (2019), Human emotional vocalizations can develop in the absence of auditory learning., *Emotion*.
- Scherer, K., A. Schorr, and T. Johnstone (2001), *Appraisal Processes in Emotion: Theory, Methods, Research*, Series in Affective Science, Oxford University Press.
- Selye, H. (1984), *The Stress of Life*, McGraw-Hill Education.
- Shuman, V., D. Sander, and K. Scherer (2013), Levels of valence, *Frontiers in Psychology*, 4, Article 261.
- Steimer, T. (2002), The biology of fear- and anxiety-related behaviors, *Dialogues in clinical neuroscience*, 4, pp. 231–49.
- Stein, N. L., B. Leventhal, and T. Trabasso (1990), *Psychological and biological approaches to emotion*, L. Erlbaum Associates Hillsdale, N.J.
- Stevens, C., and D. Bavelier (2012), The role of selective attention on academic foundations: A cognitive neuroscience perspective, *Developmental cognitive neuroscience*, 2 Suppl 1, pp. 30–48.
- Sullivan, M., and M. Lewis (2003), Emotional expressions of young infants and children, *Infants & Young Children*, 16, pp. 120–142.

- Sun, R., N. R. Wilson, and M. Lynch (2015), Emotion: A unified mechanistic interpretation from a cognitive architecture, *Cognitive Computation*, 8, pp. 1–14.
- Tao, J., and T. Tan (2005), Affective computing: A review, in *Affective Computing and Intelligent Interaction*, edited by J. Tao, T. Tan, and R. W. Picard, pp. 981–995, Springer Berlin Heidelberg, Berlin, Heidelberg.
- Thompson, R. (2001), *Infancy and Childhood: Emotional Development*, pp. 7382–7387.
- Vallverdu, J., J. Vallverdu, and D. Casacuberta (2009), *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, IGI Global, USA.
- Wang, P. (1995), Non-axiomatic reasoning system: Exploring the essence of intelligence, Ph.D. thesis, USA.
- Wang, P. (2006), *Rigid Flexibility: The Logic of Intelligence*, Springer, Dordrecht.
- Wang, P. (2013), *Non-Axiomatic Logic: A Model of Intelligent Reasoning*, World Scientific, Singapore.
- Wang, P., M. Talanov, and P. Hammer (2016), The emotional mechanisms in NARS, in *Artificial General Intelligence - 9th International Conference, Lecture Notes in Computer Science*, vol. 9782, pp. 150–159, Springer.
- Wang, P., X. Li, and P. Hammer (2018), Self in NARS, an AGI System, *Frontiers in Robotics and AI*, 5, Article 20.
- Wang, P., P. Hammer, P. Isaev, and X. Li (2020), The conceptual design of opennars 3.1.0, *Tech. rep.*, Temple University, Philadelphia, United States.
- Yerkes, R. M., and J. D. Dodson (1908), The relation of strength of stimulus to rapidity of habit-formation, *Journal of Comparative Neurology and Psychology*, 18(5), pp. 459–482.
- Zhan, J., X. Wu, J. Fan, J. Guo, J. Zhou, J. Ren, C. Liu, and J. Luo (2017), Regulating anger under stress via cognitive reappraisal and sadness, *Frontiers in Psychology*, 8, Article 1372.

APPENDIX

NARSESE EXAMPLE FOR CHAPTER 6

=====Happiness=====

//Meaning of the statement: If something is wanted by SELF,
//and SELF's belief agrees with the case, SELF feels Happy

//1. #1 is a dependent variable which represents a certain
// unspecified term under a given restriction. It can be
// either an object or an event

//2. (^want, {SELF}, #1, TRUE) represents a mental operation
// means something is desired by SELF; TRUE indicates the
// truth value of this mental operation, where #1 is desired,
// otherwise, use FALSE

//3. (^believe, {SELF}, #1,TRUE) means SELF's belief agree
// with #1, if #1 represents an event, it indicates that
// #1 has already happened.

//4. (^feel, {SELF}, happy) implements feel operator and

```

// indicates the feeling of SELF being happy

//5. && is a term connector, it connects the follow
// term by meaning ‘and’

Input: <(&&, (^want, {SELF}, #1, TRUE), (^believe, {SELF},
#1,TRUE)) =|> (^feel, {SELF}, happy)>.

//SELF has a goal which is not being hurt, ‘--’ is the negation
//of the statement

Input: (--,<{SELF} --> hurt>)!

//SELF is not getting hurt, :|: represents the tense ‘present’
//means SELF is not getting hurt right now

Input: (--,<{SELF} --> hurt>). :|:

//What do you feel?
//This statement is a question, and it corresponding to
//(^feel, {SELF}, happy) where ‘?what’ at the position of the
//emotion

Input: (^feel,{SELF},?what)?

//SELF feels Happy, the reason why it feels happy is because
//SELF doesn’t want to get hurt (generated by goal), and SELF
//is not getting hurt (generated by belief).

```

Answer: (^feel,{SELF}, happy).

=====Fear=====

//If something is wanted by SELF, and SELF anticipates the
//opposite to happen, SELF feels fear

Input:<(&&, (^want, {SELF}, #1, FALSE), (^anticipate, {SELF},
#1)) =|> (^feel, {SELF}, fear)>.

//At the same time when SELF feels fear, it generate an
//motivation which to run away, run is also an operator in NARS

Input: <(^feel,{SELF}, fear) =|> <(*, {SELF},
<(*, {SELF}) --> ^run>) --> ^want>>.

//SELF doesn't want to be hurt

Input: (--,<{SELF} --> hurt>)!

//If wolf is getting close to SELF, SELF will get hurt
//&/ is another term connector representing the relation between
//two terms is ‘‘and’’, also the latter happens after the former.
//42 represents inference steps, it means, when wolf start
//getting close to SELF, after 42 steps, the SELF will get
//hurt. The number is not fixed, it can be any integer.

Input: <(&/,<(*, {SELF}, wolf) --> close_to>,+42) =/>
<{SELF} --> [hurt]>>.

//Wolf is getting close to self

Input: <(*, {SELF}, wolf) --> close_to>. :|:

//Result: SELF takes the action run, based on the knowledge

//where SELF runs when it feels fear, SELF also feels the emotion

//fear

EXECUTE (^run,{SELF})

LINKS TO NARS PROJECT

Link of OpenNARS 3.1.1:

https://github.com/opennars/opennars_core/releases/tag/v3.1.1

Link of OpenNARS Project:

<https://github.com/opennars/opennars>