Conceptual and Normative Issues of Memory Enhancement

Inauguraldissertation

zur Erlangung des Akademischen Grades

eines Dr. phil.,

vorgelegt dem Fachbereich 05 – Philosophie und Philologie

der Johannes Gutenberg-Universität Mainz

von

Ying-Tung Lin

aus Taipei, Taiwan

2015

Tag des Prüfungskolloquiums: 15 Juli 2014

Table of Contents

List of Tables iv

List of Figures v

Introduction 1
0.1 The Epistemic Goals 2
0.2 The Argumentative Goals 6
0.3 Summary of Chapters 7

Chapter 1 Catalog: Normative Issues about Memory Enhancement and Modification 12
1.0 Introduction 12
1.1 Issues at the Personal Level 13
1.2 Issues at the Social Level 21
1.3 Summary 23

Chapter 2 Conceptual Tools I: Memory 25
2.0 Introduction 25
2.1 What Is Memory? 26
2.2 The Representational Theory of Memory 41
2.3 Constructive Memory 52
2.4 Summary 60

Chapter 3 Conceptual Tools II: Selfhood, Personhood, and Identity 63
3.0 Introduction 64
3.1 Self and Self-Consciousness 64
3.2 The Autobiographical Self-Model and Memory 72
3.3 What I Am 90
3.4 Transtemporal Identity 103
3.5 Summary 119

Chapter 4 Conceptual Tools III: Enhancement 121
4.0 Introduction 121
4.1 Utilitarianism and Suffering 123
4.2 The Conceptual Issues of Enhancement 130
4.3 The Concept of Health 137
4.4 The Phenomenological Account of Health and Enhancement 144
4.5 Summary 149

Chapter 5 Memory Interventions: The Current Situation 151
5.0 Introduction 151
5.1 Interventions at the Molecular Level 152
5.2 Brain Stimulation 160
5.3 Physical and Mental Exercises 163
5.4 The Classification of Memory Interventions 164
5.5 Summary 167

Chapter 6 A Fresh Look at the Concept of Memory Enhancement 169
6.0 Introduction 169
6.1 A Better Memory? The Conceptual Issues 170
6.2 Suffering and Memory Malfunction 173
6.3 Self-Interest and Memory Enhancement 180
6.4 The Phenomenological Account of Memory Enhancement 187
6.5 Summary 189

Chapter 7 Authenticity: The Worry of the Loss of the Self and Identity 191
7.0 Introduction 191
7.1 The Concern of Authenticity in the Cognitive Enhancement Debate 192
7.2 Authenticity as Self-Discovery 194
7.3 Authenticity as Self-Creation 204
7.4 The Subjective and Objective Senses of Authenticity 211
7.5 Summary 212

Chapter 8 Memory Intervention: A Means or a Threat to Authenticity? 214
8.0 Introduction 214
8.1 The Framework of Authenticity 215
8.2 The Relationship between Authenticity and Memory Manipulation 220
8.3 The Moral Value of Authenticity 222
8.4 Authenticity and Memory Enhancement 224
8.5 Summary 226

Chapter 9 Conclusion 228
9.0 Introduction 228
9.1 Summary of the Discussions 229
9.2 The Phenomenological Account of Memory Enhancement 234
9.3 Memory Enhancement and the Issue of Authenticity 237
9.4 Open Questions for Future Studies 241

Epilogue 244

References 246

List of Tables

Table 1. *The Differences between Episodic and Semantic Memory.* 37

Table 2. *Different Conceptions of Normal.* 139

List of Figures

Figure 1. A classification of memory types..... 32

Figure 2. A simplified sketch of memory processing..... 33

Figure 3. Personhood and the self-models. 101

Figure 4. The molecular mechanisms of early and late long-term potentiation (LTP). . 155

Figure 5. Categorization of memory interventions..... 166

Figure 6. The model of memory treatment and enhancement. 170

Figure 7. The phenomenological model of treatment and enhancement..... 187

Introduction

0.1 The Epistemic Goals
0.1.1 Conceptual Issues 1: "Memory"
0.1.2 Conceptual Issues 2: "Enhancement"
0.1.3 Conceptual Issues 3: "Authenticity"
0.1.4 Conceptual Issues 4: "Selfhood", "Personhood", and "Identity"
0.1.4 Conceptual Issues 5: "Autobiographical Self-Model"
0.2 The Argumentative Goals
0.2.1 The Phenomenological Account of Memory Enhancement
0.2.2 The Issue of Authenticity in Memory Enhancement
0.3 Summary of Chapters

With the growing understanding of human brains and cognition, and the development of neurotechnology, we have ever-increasing power to intervene in the human mind by modifying the neural and bodily processing. Nowadays, new knowledge and technology are not only used to treat patients, but are also applied to the healthy for enhancement purposes. Pharmaceutical drugs are one of the examples. For instance, some of the most prevalent cognitive enhancers are the psychostimulants designed for attention-deficit hyperactivity disorder (ADHD), e.g., Ritalin® and Adderall®. These drugs have been taken by American university students to improve their concentration levels as well as by fighter pilots to enhance alertness (McCabe, Knight, Teter, & Wechsler, 2005; Teter, McCabe, LaGrange, Cranford, & Boyd, 2006). Another kind of cognitive enhancers is anti-depressant like Prozac® or Paxil®. It has been reported that these mood-enhancing agents are used by one in eight American adults, while only around 3–5% males and 8–10% females are actually diagnosed with depression (Coenen et al., 2009; President's Council on Bioethics, 2003).

In addition to attention and mood, memory is another cognitive function targeted for enhancement. It is one of the most important core cognitive faculties for human beings. It not only provides the capacity for learning and retaining new information to increase our behavioral flexibility, but by supporting a mental autobiography, also allows us to have self-knowledge concerning who we are and how we are related to the external world and other social beings. That is, it allows us to have a feeling of continuity and identity: Phenomenally, we experience ourselves embedded in the (social) environment and living in the present with a

personal history in the past and expectations and plans for the future. Furthermore, recent studies (De Brigard, Addis, Ford, Schacter, & Giovanello, 2013; Schacter & Addis, 2007a; 2007b; see §2.3) have shown that the constructive nature of memory allows us to simulate possible scenarios, which support the capacities of decision making and planning.

Current memory modification technologies (MMTs; Liao & Sandberg, 2008) aim at either strengthening or weakening one's capacity of memorizing. Psychoactive substances with the former function include donepezil (e.g., Aricept®), which is prescribed as a treatment for Alzheimer's disease, and Ginkgo biloba. Beta-blockers, on the other hand, used in cases of post-traumatic stress disorder (PTSD), can diminish the emotional content of memories. It is believed that in addition to pharmaceuticals, more techniques aiming at altering memory are under development and will be available soon. (Candidates of memory enhancers will be reviewed in §5.)

In view of the possible prevalence of cognitive enhancers in the near future, the ethical issues surrounding cognitive enhancement (CE) have been a subject of debate among academics. In discussion of human enhancement, the controversial issue has been whether we should utilize technologies to enhance human capacity. There are two main competing positions on this issue: Transhumanists argue that the human species is in an early stage of evolution and will at some point become post-human, bearing much greater capacities through the use of science, technology, and other possible means (Bostrom, 2003; Schneider, 2009). On the contrary, bioconservativists worry that enhancement technologies will undermine human virtues, which are central to human beings. This dissertation aims to touch the debate in the context of memory enhancement. I will focus on two normative issues surrounding memory enhancement: (1) the distinction between memory treatment and enhancement—how is memory enhancement distinguished from memory treatment among memory intervention?—and (2) the issue of authenticity in the context of memory enhancement—how does the issue of authenticity concern the moral permissibility of memory enhancement?

0.1 The Epistemic Goals

With a view to providing a constructive proposal to these normative issues, the following concepts need to be delineated. These include the concepts of memory and enhancement—for the argument for a distinction between memory treatment

and enhancement—as well as the notion of authenticity—to explore how the issue of authenticity relates to memory intervention or enhancement. I will examine different conceptions of authenticity, the related concepts of selfhood and personal identity, and finally the idea of an autobiographical self-model (ASM), which I develop to provide a framework for considering the issue of authenticity. The following three sections (§2-4) aim to build the conceptual foundation.

0.1.1 Conceptual Issues 1: "Memory"

The definition of "memory enhancement" relies on the definitions of "memory" and "enhancement". How is "memory" defined? I define general memory processing as short- or long-term changes in neurocognitive processing which directly results from a corresponding experience. However, the term "memory" is used to refer to a variety of things as well as different kinds of memory; how are different concepts of memory distinguished and how are they related to one another? These concepts are reviewed and clarified to avoid later confusion.

What is the nature of memory? Based on the representational theory, memory is regarded as a special kind of mental representation—mental simulation. The representational theory of memory, which posits a tripartite relationship between representation, representandum, and representatum is adopted in this dissertation (Metzinger, 2004). However, how can the representational theory characterize differences between representations involved in perception and those involved in memory? The distinction between representation and simulation accounts for the conceptual and functional differences between perception and recollection.

I will also discuss the function of memory. The function of memory will play an important role in determining the distinction between memory treatment and enhancement. A new understanding of the function of memory emerges from studies of misremembering and constructive memory. Empirical studies support the idea that memory exists in order to increase an individual's behavioral flexibility, rather than to re-present past information correctly (Suddendorf & Corballis, 2007). The increase of behavioral flexibility promotes the adaptability to the changing of environment of the individual organism.

0.1.2 Conceptual Issues 2: "Enhancement"

The term enhancement is used in bioethics "to characterize interventions designed to improve human form or functioning beyond what is necessary to sustain or

restore good health" (Juengst, 1998, p. 29). The concept of "enhancement" presupposes (1) becoming better, i.e., improvement and (2) a distinction between treatment and enhancement.

Concerning the concept of "memory improvement", we ask: What is a better memory, or better memory function? How do we judge whether one memory function or state is better than another? It is argued that one cannot attribute normative properties to a memory function or state simply by considering the properties of the memory; instead, one has to investigate how memory contributes to an individual's welfare—either in terms of alleviating individual suffering or by promoting the individual's self-interest.

The distinction between enhancement and treatment relies on the distinctions between the concepts of health and ill-health and the concepts of normal and abnormal. Different conceptions of normality and theories of health are reviewed, and, based on the negative utilitarianism, which is adopted as the practical normative theory of this dissertation, the phenomenological account of health (PAH) is introduced. According to the PAH, ill-health is distinguished from health by a demarcation criterion: the existence of suffering. The phenomenological account of the distinction between treatment and enhancement is thus determined by the PAH.

0.1.3 Conceptual Issues 3: "Authenticity"

"Authenticity" is regarded as a moral ideal by both critics and proponents of CE. The former claim that CE threatens our authentic lives, while the latter argue that CE can be a means to living authentically. This discrepancy results from differences in their understanding of both the concept of authenticity, the ideas of the self and identity they endorse, and the kind of cognitive enhancers they each have in mind.

To explore this issue, different understandings of the concepts are reviewed: Authenticity as self-discovery concerns our conformity to our own internal framework; by contrast, authenticity as self-creation emphasizes one's identification with an action. These two conceptions have built on two distinct ideas of the self: The former holds that there exist some pre-given, psychological characteristics which constitute one's true self and are deterministic of one's identity; while the latter deny the existence of any pre-given self. Consequently, while both value the ideal of authenticity, these two positions have different concerns when it comes to the utilization of CE.

0.1.4 Conceptual Issues 4: "Selfhood", "Personhood", and "Identity"

In order to look into the conceptual issues of authenticity, some related concepts, including "self" and "personal identity" are considered. First, the concept of authenticity understood as "being true to one's self" presupposes that there exists such a thing as the self. However, this concept may be conceptually problematic as it is based on a kind of ontological realism of the self. This dissertation is based on the self-model theory (SMT) that accounts for our phenomenal experience of being a self or a subject through representational systems and functional constraints (Metzinger, 2004): What we experience as a self is the content of a phenomenal self-model.

Second, one of the concerns relating to authenticity in the CE debate results from the worry that it may lead to loss of one's identity. As such, different concepts of identity are discussed, including transtemporal human identity and transtemporal personal identity. The latter relies on the concept of a person, which I consider the property of the whole system that emerges when a human being is equipped with the mental capacities required for interaction in its moral community.

0.1.5 Conceptual Issues 5: "Autobiographical Self-Model"

The concept of authenticity, built upon the existence of the self, is conceptually problematic. However, both critics and proponents of CE have been concerned with the issue of authenticity in terms of one's mental autobiography. Based on the SMT, the concept of an ASM is developed to account for this concern. An ASM refers to a collection of mental simulations which result in one's mental autobiography constituted by one's past and potential future states and their relations to the current state.

Three functional constraints for a self-model to be considered an ASM are related to the issues of authenticity: synchronic coherence, diachronic coherence, and global veridicality. Synchronic coherence refers to the consistency of the contents of simulata in the ASM constructed at a time; diachronic coherence refers to consistency between the contents of ASMs constructed at different times; and global veridicality refers to the degree to which an ASM corresponds to past experience and events. The examination of these properties of an ASM is later used to consider the concerns involved in the issue of authenticity.

0.2 The Argumentative Goals

Two normative issues of memory enhancement are considered in this dissertation: The first considers the distinction between memory treatment and enhancement; the second considers the issue of authenticity in the context of memory enhancement and memory manipulation.

0.2.1 The Phenomenological Account of Memory Enhancement

How is memory enhancement distinguished from memory treatment? How can we determine if one memory function or state is better than another is? I argue for the phenomenological account of memory enhancement. This is based on the phenomenological account of health and the health-based account of enhancement: According to the former, a state of health or ill-health is determined by the existence of suffering; then the distinction between treatment and enhancement is made depending on whether the system it addresses is healthy or not. Therefore, suffering plays a critical role in determining enhancement.

Applied to the issue of memory enhancement, memory treatments are, under standard circumstances, interventions that address malfunctions of memory that result in an individual's suffering or potential suffering, and from which the subject has no ability to independently escape. On the other hand, memory enhancements are, under standard circumstances and without any unwilling suffering or potential suffering resulting from the alteration of memory functions, interventions that aim to manipulate memory function based on the self-interests of the individual.

Two points are critical for determining memory enhancement according to the phenomenological account. First, one has to be free from suffering that might result from memory malfunction. Therefore, it will be crucial to investigate the relationship between suffering and malfunction of memory. Second, one's self-interest determines what is a better memory function or state when one is free from suffering. That is, one's identification and preference determines what can be considered memory enhancement, and this may differ from individual to individual.

0.2.2 The Issue of Authenticity in Memory Enhancement

The second issue with which this dissertation is concerned is the issue of authenticity. How do memory enhancement and modification relate to the issue of authenticity? Is memory enhancement a means or a threat to an authentic life?

First, I argue that understanding the concept of authenticity as being true to oneself is conceptually problematic. The concept of authenticity is understood as being true to one's self; however, this conception depends on the presupposition that there exists something that is "the self" that provides contents to which we can conform. So far, empirical studies have provided no evidence to support ontological realism of the self.

Second, I argue that the debate between critics and proponents of CE can be understood as concerns with (psychological) identity, autonomy, and truthfulness, which can further be illustrated with three constraints of ASM—diachronic coherence, synchronic coherence, and global veridicality. This way of considering the characteristics of ASM allows us to consider how the permissibility of memory intervention is affected by these concerns.

Third, the above concerns and memory interventions are engaged in a two-way relationship: On the one hand, the concerns confining the permissibility of memory interventions can be understood in terms of functional constraints for an ASM. On the other hand, memory interventions by altering the content of the self-model can affect how the system has to be in order to satisfy the functional constraints for an ASM; that is, memory intervention can alter the criterion for the concerns.

Last, the value of the concerns of truthfulness, autonomy, and (psychological) identity will be examined through investigating the relationship between the constraint satisfactions and suffering. Only autonomy is necessary and should be pursued without further consideration, as synchronic incoherence is sufficient for suffering. Truthfulness or (psychological) identity, however, is not sufficient for confining the permissibility of memory intervention or memory treatment.

0.3 Summary of Chapters

The dissertation is divided into two parts: The first part (§2–§4) consists of conceptual analysis of the concepts required for the normative considerations; the second (§5–§8) is a synthesis of two normative issues: the distinction between memory treatment and enhancement and the debate of authenticity in the context of memory enhancement.

Chapter 1—Catalog: Ethical Issues about Memory Enhancement

It is believed that the development of memory modification technologies will continue in the near future. Thus, it is crucial that we identify relevant moral issues related to memory modification in advance. The first chapter lists ethical issues concerning memory enhancement and interventions. These includes the issue of safety, the distinction between memory treatment and enhancement, concerns of authenticity, cognitive liberty and autonomy, truthfulness, moral enhancement, indirect coercion, the right and responsibility to remember, and applications in special subjects and special situations. Only the distinction between memory treatment and enhancement and the issue of authenticity are focused on in this dissertation, but other answers are implied. The conceptual tools developed in the first part of the dissertation can be utilized for consideration of other issues.

Chapter 2—Conceptual Tools I: Memory

This chapter aims to clarify the cognitive system in question. What different concepts and different kinds of memory do we understand? What is the nature of memory that makes it possible for us to recall experiencing something in the past? What is the function of memory? Is misremembering a memory malfunction? This chapter first reviews different ways in which the term memory is used and classified. The concepts representing different kinds of memory—"working memory", "episodic memory", "semantic memory", and "autobiographical memory", and their relationships and distinctions are reviewed and discussed (§2.1). Then, a version of representational theory of memory is adopted: The characteristics of memory are illustrated through distinctions between concepts of representation and simulation, between concepts of simulation and self-simulation, and between concepts of mental and phenomenal (self-)simulation (§2.2). Last, a newly developed view of memory function and malfunction is introduced. This results from empirical studies on the constructive nature of memory and a new way of looking at misremembering (§2.3).

Chapter 3—Conceptual Tools II: Selfhood, Person, and Identity

This chapter aims to delineate the concepts of "the self", "person", and "identity". First, what is the concept of the self? What are the target phenomena and the questions to be considered? How do we account for the phenomenal self and for subjectivity? What is the nature of these phenomena? The SMT provides a

comprehensive representational and functional theory to account for self-related phenomena: Self-models, together with functional constraints, serve to explicate different forms of self-consciousness, and give a picture of the nature of the phenomenal self and subjectivity (§3.1). Second, how are different kinds of memory related to the self-model? How do we account for our mental autobiography? What is personality? I introduce the concept of an ASM and consider its relation to memory and personality (§3.2). Then I ask, what essentially am I? What does the concept of a person mean? Why are these questions morally significant? (§3.3) To understand what I essentially am, animalism and the psychological approach are discussed. I argue for an alternative nature of "I" and a normative concept of the person. Last, I turn to the concept of identity. What is our concept of identity? What are the defining criteria for an identity? I differentiate the concepts of transtemporal human identity and transtemporal personal identity. The identity criteria of these concepts rely respectively on concepts of human organism and person (§3.4).

Chapter 4—Conceptual Tools III: Enhancement

The last concept I consider is "enhancement". In this chapter, I first introduce negative utilitarianism, a default practical normative theory of this dissertation, and the relationship between suffering and self-interests (§4.1). I then examine the conceptual issues required for delineation of "enhancement", including the concepts of health, disease, illness, and normality (§4.2). I look at two rival accounts of health: the biostatistical theory (Boorse, 1975, 1977) and the holistic theory (Nordenfelt, 1993, 2001, 2007) (§4.3). I introduce a revised version of the holistic theory, namely the phenomenological account of health. Based on this account, I introduce the phenomenological concept of enhancement (§4.4).

Chapter 5—Memory Interventions: The Current Situation

My examination of the issue of memory enhancement starts with an overview of the current situation regarding memory intervention and memory enhancement. I review different kinds of candidates for memory enhancers, including pharmaceutical interventions, genetic engineering, herbs and nutrition, brain stimulations, and physical and mental exercises (§5.1–§5.3). Then, the possible effects of memory intervention are classified. This classification can be used to consider how a targeted memory intervention affects the constraint satisfaction for being considered an ASM (§5.4).

Chapter 6—A Fresh Look at Memory Enhancement

To apply the phenomenological concept of enhancement on "memory enhancement", two additional conceptual issues have to be considered: (1) the function and malfunction of memory, and the relationship between memory malfunction and suffering; (2) the criteria for a better memory, which will determine memory enhancement. The first issue is based on a new way of understanding misremembering, which considers the function of memory as increasing the behavioral flexibility of the organism. Different ways in which memory malfunction leads to suffering are discussed (§6.2). The second issue involves concerns of paternalism and autonomy. Without the existence of memory-related suffering, a better memory is determined by the self-interests of the individual (§6.3).

Chapter 7—Authenticity: The Worry of the Loss of the Self and Identity

§7 and §8 deal with the ethical issue of authenticity in the context of memory enhancement. §7 considers the concept of authenticity. I respectively review the positions adopted by critics and proponents of CE. These relate to authenticity as self-discovery (§7.2) and authenticity as self-creation (§7.3), background assumptions of "the self" and "identity", and attitudes toward CE.

Chapter 8—Memory Intervention: A Means or a Threat to Authenticity?

After examining the conceptions of authenticity endorsed by critics and proponents of CE, this chapter aims to characterize their concerns by looking at the functional constraints for a self-model to become an ASM; this will allow us to examine the relationship between authenticity and memory intervention and enhancement. First, three concerns are at issue: truthfulness, autonomy, and identity. These can respectively be characterized by the functional constraints of global veridicality, synchronic coherence, and diachronic coherence. The characterization of these concerns as constraint satisfaction provides a framework in which to examine how the ASM can be affected, and by considering the relationship between constraint satisfaction and suffering, the value of the concerns is reconsidered. Last, the two-way relationship between memory intervention and concerns involved in authenticity is investigated.

Chapter 9—Conclusion

The last chapter summarizes the discussions provided in the dissertation. I first review the results of the discussion in each chapter. These results might form a theoretical foundation for other normative issues concerning CE or memory enhancement and intervention. I then summarize the claims and arguments for the phenomenological account of memory enhancement and the issue of authenticity in memory enhancement.

Chapter 1

Catalog: Normative Issues about Memory Enhancement and Modification

1.0 Introduction
1.1 Issues at the Personal Level
1.1.1 The Distinction between Memory Treatment and Enhancement
1.1.2 Safety
1.1.3 Authenticity
1.1.4 Identity
1.1.5 Cognitive Liberty and Autonomy
1.1.6 Truthfulness
1.1.7 Moral Enhancement
1.1.8 Special Subjects and Situations
1.2 Issues at the Social Level
1.2.1 Indirect Coercion
1.2.2 The Responsibility to Remember and the Right to Forget
1.3 Summary

1.0 Introduction

This chapter will provide a list of normative issues surrounding memory enhancement and modification. The first section considers issues at the personal level, that is, the consideration of these issues generally concerns the individual in question alone. They include the issue of safety, the distinction between memory treatment and enhancement, the issue of authenticity, cognitive liberty and autonomy, truthfulness, moral enhancement, and special subjects (e.g., children, old people, patients, animals, and future subjects) and situations (e.g., conditions of illness, military service, and experiment or research). The second section concerns issues at the social level, including the concern of indirect coercion, the responsibility to remember, and the right to forget.

However, the distinction between personal and social levels is merely a general categorization. The level at which these issues are relevant may depend on the account of the issues offered. For instance, when it comes to the distinction

between treatment and enhancement, a distinction can be made in terms of properties of the individual (health or disease) alone, whereas some have claimed that the distinction should also be based upon distributive justice, which pertain to how an intervention affects not just one but many individuals.

1.1 Issues at the Personal Level

1.1.1 The Distinction between Memory Treatment and Enhancement

How can memory treatment and enhancement be distinguished? This distinction is regarded as the basis of discussion of other normative considerations regarding memory enhancement, which are listed in the remainder of the chapter. The distinction may seem clear at the first glance. It presupposes a direction of betterment—the criteria indicating what a better memory state or function is—and a distinction between types of interventions—the criteria indicating two kinds of interventions. We can generally define memory treatment and enhancement as follows: The former is the use of memory intervening technologies to treat individuals with memory malfunction, memory disorder, or abnormal memory capacity, in order to reach a healthy or normal capacity of memory. By contrast, the latter aims to improve the normal working of memory functions of a healthy individual.

However, these definitions presuppose a clear understanding of descriptive concepts, e.g., "memory" and "memory function", of normative concepts, e.g., "normal and abnormal", and "improvement", and controversial concepts, e.g., "health, disease, illness, and disorder".[1] That is, in order to develop clear concepts of memory treatment and enhancement, an examination of these concepts is required. The relevant questions involved in developing a distinction include the following:

- How is memory enhancement distinguished from memory treatment? What is the demarcation criterion?
- What are memory function and malfunction? What is considered "normal" or "better" memory function?

[1] As I will discuss in §4, whether the concept of health is normative or descriptive is controversial: The objectivists hold that the concepts of health in contrast to the concept of disease is purely objective, whereas the constructivists argue that the concepts are normative in nature.

- Is there a universal criterion for "better" memory or memory enhancement, which is applicable to all individuals?
- What is the function of the treatment-enhancement distinction in memory interventions?

Issues surrounding the distinction between treatment and enhancement will be discussed in §4, and the application of the distinction to memory intervention will be dealt with in §6.

1.1.2 Safety

Safety issues surrounding memory intervention arise due to our ignorance of the benefits, side effects, and risks that could result from memory interventions as well as our conceptions of "side-effect" and "risk". Our lack of knowledge originates in the difficulty of assessing the effect of memory interventions. The difficulty comes from (1) a lack of longitudinal studies on memory enhancement, and (2) a lack of empirical studies on healthy volunteers.

In addition, the complexity of the mechanisms of memory also leads to a further complication, namely *designing* memory interventions (Glannon, 2006, p. 77). First, neural correlations are distributed across different region of the brain (McClelland & Rumelhart, 1985), and as such memory functions cannot be modified by simply altering one or a localized group of synaptic connections. In order to attain the desired result in memory intervention, more empirical studies of underlying mechanisms and the minimal set of connections are required. Second, memory systems and other cognitive systems, e.g., attention, emotion, language, reasoning, and so on, can influence each other. Thus, by modifying one's memory capacity, one may influence other cognitive functions; for instance, dampening one's episodic memory can diminish one's capacity for decision-making. On the other hand, one's memory function can also be intervened with indirectly by modifying other cognitive faculties, for example, increasing one's attention can indirectly increase one's capacity for memorizing.

The concept of memory enhancement is determined through an assessment of memory intervention, while the necessity of the assessment relies on the moral permissibility of memory enhancement (Metzinger & Hildt, 2011). First, whether a memory intervention is considered enhancement relies on an assessment of its effects on each individual (, which may differ according to one's cognitive baseline or level of cognitive function before the intervention). Although this relies on the

account of memory enhancement one endorses, its efficacy will determine the possibility of its being considered an enhancement. For instance, according to the phenomenological account of memory enhancement that I examine in §6, if a memory intervention leads to an increase in memory-related suffering, it cannot be considered memory enhancement for the subject in question. Second, the call for longitudinal studies on healthy subjects relies on the idea that memory enhancement is permissible. If conceptually memory enhancement were impermissible, there would be no need for such an assessment. However, there is no normative argument that is sufficient to prove impermissibility, and, in addition, the normative claim should be based on empirical studies of the effects of memory interventions. Therefore, studies of the effect of memory interventions, either treatment or enhancement, are required and will contribute to more relevant discussions of the conceptual and normative issues of memory modification.

1.1.3 Authenticity

"Authenticity"—understood as "being true to oneself"—is considered a moral ideal by both critics and proponents of cognitive enhancement (CE). Understanding the concept differently and endorsing distinct ideas of the "self", they hold opposing views of whether CE is a threat or a means to an authentic life. Memory alteration is especially related to the concern of authenticity.

The issue of authenticity in the context of memory enhancement and intervention concerns whether intervening in one's memory makes one's life more or less authentic. As we will see, debate on the issue mainly focuses on self-knowledge, which is supported by one's memory processing. Accordingly, memory interventions that alter underlying processes can modify one's self-conception. As such, the issue of authenticity lies in the question of what kind of alteration in self-conception would lead to an authentic or inauthentic life:

- What does the concept of authenticity mean? What theory of the "self" is endorsed?
- For critics of CE, what does it mean to lose the self? Will memory enhancement through biotechnology result in loss of the self? On the other hand, for proponents of CE, what does it mean to enhance the self? What is a better self? Can we enhance the self by manipulating memory?
- Is authenticity a moral ideal that is worth pursuing? Can it be regarded as a sufficient criterion for considering the permissibility of an intervention?

The issue of authenticity in the context of memory intervention or enhancement is one of the main normative issues with which I will be concerned. In §7, I review and analyze different conceptions of authenticity and accounts of the "self" respectively endorsed. In §8, I confront the issue in relation to memory intervention and investigate the relation between authenticity and memory modification.

1.1.4 Identity

The general issue of personal identity[2] concerns the transtemporal identity relation of individuals at different times. The debate over the criterion for the transtemporal identity relation has not been resolved, but can be divided into the biological and the psychological approach: The former states that the identity relation is determined by the sameness of biological properties; the latter claims that it requires the continuity of psychological characteristics.

Does memory intervention affect transtemporal identity? This question will be answered differently depending on the concept of personal identity one endorses. If one adopts a biological approach to identity, memory intervention cannot affect one's identity. Nevertheless, where the psychological approach is endorsed, memory intervention can affect identity, since the continuity of memory is a criterion that allows the identity relation to hold. Finally, yet importantly, in order to argue that memory modification is unethical because it is dangerous—in the sense that it might leads to loss of identity—one must also account for *why* it would be morally problematic to lose this identity. Without this link, one cannot argue that memory intervention is dangerous because of loss of identity.

- What is the criterion for personal identity—or the criterion for the individuals at different times to be identical?
- Can memory be modified or enhanced without changing the identity of a person?
- Does the issue of identity matter? If it does, why is it morally problematic to alter one's identity?

[2] I use "the general issue of personal identity" to refer to the traditional philosophical issue of personal identity. Later in §3, I differentiate the concepts of "transtemporal human identity" and "transtemporal personal identity".

1.1.5 Cognitive Liberty and Autonomy

"Cognitive liberty" refers to one's right to maintain autonomy over one's mind (Bublitz, 2013; Sententia, 2004). This idea suggests that without violating other strong moral concerns (e.g., harming other people), we have the fundamental right to self-govern our mind and brain, including utilizing brain intervention technologies to modify our own cognitive function or to manipulate our mental contents. Cognitive liberty is considered an argument in favor of CE.

Conceptually, cognitive liberty implies that we are free to decide not only which mental states and capacities we have but also our mental contents. Do we have the right to memorize or forget whatever we want? We seem to have such right as long as we consider such action autonomously: We are fully aware of our self-interests and possible outcome of the intervention, and to be autonomous, we prefer and identify the action and possible result of the intervention. However, by definition, for one to memorize or recall, instead of imagining, one has to believe that something has really happened—that it is the truth or exists in the past—even though it has never happened or happened the way it is recalled. Consider a person who modifies her memories of childhood (or, more dramatically, all of her memories about her past) into an experience she has longed for, based on this right. After the memory modification, this person truly believes that the new memories result from real experience, rather than from memory design. Can this person still be considered autonomous, after losing a large and critical part of information concerning the modification? Does cognitive liberty presuppose sufficient information?

If cognitive liberty requires the subject to be fully informed, we have to ask: Is it possible, conceptually, for one to be fully informed? Otherwise, how well-informed is sufficient? What is peculiar to cognitive intervention, especially interventions in memory and emotion, is its effect on personality, value, and preference, which are characterized by the autobiographical self-model (ASM; see §3). Any transition, mild or severe, is likely to result in changing one's attitude toward something, including the intervention which causes the conversion. Before the intervention, provided we have enough relevant information, we can imagine what it will be like afterwards. Nevertheless, such mental simulation is limited, because it is based on a different ASM. Therefore, one can never be completely informed, and this is notably evident in the case of memory intervention.

The issue of cognitive liberty is closely related to other normative issues. For instance, coercion can result in loss of one's cognitive liberty; thus, preventing

coercion promotes autonomy. On the other hand, other issues are in conflict with reinforcing cognitive liberty. For example, truthfulness and responsibility of remembering seem to limit cognitive liberty. These two issues will be introduced later in this chapter.

- Do we have the fundamental right to autonomy over our memory? Can one always remain autonomous after memory modification?
- Does cognitive autonomy presuppose one being well-informed? What are the conceptual and practical difficulties of sufficiently informing the subject? How well-informed is sufficient for the subject's autonomy?
- Is it possible to be an autonomous agent without any external influences?
- How can we resolve the tension between the normative concerns of cognitive liberty and truthfulness or responsibility to remember?

1.1.6 Truthfulness

Remembering or misremembering presupposes that the subject believes that the content of a memory corresponds to the reality; that is, what one recalls must have happened. Without this belief, one is imagining, instead of remembering, and will lose the phenomenal experience that accompanies the memory—the sense of pastness and familiarity. However, a feeling of accuracy does not imply that the content of a recollection is real. Different forms of misremembering are pervasive in human beings (Schacter, 2001), and result from the constructive nature of memory (see §2.3). A new account of the function of memory suggests that memory serves to increase behavioral flexibility instead of re-presenting the past faithfully (De Brigard, 2013; Suddendorf & Corballis, 2007). That is, we construct a memory during recollection, and it is likely that the content of memory is not completely veridical.

Therefore, the first question concerning the issue of truthfulness is to rethink why it is morally problematic to be untruthful. Intuitively, we consider deception wrong, unless there exist other factors to justify it.[3] However, there are cases, such as post-traumatic stress disorder (PTSD) and dissociative identity disorder (DID), which suggest that truthfulness may lead to suffering. In addition to psychological disorders, it has been suggested that self-deception be considered a kind of self-enhancement (von Hippel & Trivers, 2011). It seems that the cognitive mechanism

[3] For instance, for a consequentialist, the outcome is better with deception.

functions in such a way that stops us from having a completely veridical memory or self-knowledge. We therefore need to ask what kind of untruthfulness is morally problematic and what kind is permissible. What underlies such a distinction?

- Why is untruthfulness morally problematic?
- Is there any form of deception or self-deception that is not wrong? Under which conditions can truthfulness be sacrificed?

1.1.7 Moral Enhancement

The issue of moral enhancement concerns the possibility of enhancing one's morality by intervening in one's mind or related cognitive capacities (J. Harris, 2011) and emotion (Douglas, 2008, 2013). Moral enhancement is conceptually problematic. Defining what moral enhancement is requires clarification of what is morally better. However, this question has been the topic of hot debate and little agreement has been reached. Therefore, if we want to avoid this issue and utilize a definition of moral enhancement, we have to find a kind of alteration that is generally accepted as moral enhancement. For instance, Thomas Douglas (2008) suggests the attenuation of counter-moral emotions that can interfere with good motives or are themselves bad motives (e.g., the impulse towards violent aggression) should be accepted as moral enhancement. The second issue concerning moral enhancement is its moral permissibility. Would it be permissible to enhance our morality, if it were practically possible? Does it make a difference which subject we consider? Should we enhance a criminal's morality? If we should, what reason do we have?

Following this, to see how memory intervention can result in moral enhancement, we need to see how memory is involved in moral decision-making and behavior. As we will see later in §2.2, the ASM is constituted by underlying processes of different kinds of memory, and it is from this that one's self-interests, preferences, and values emerge. First, the capacity for moral judgment is supported by the working memory, which provides a workspace for integrating information about the current situation with information from the ASM. Second, the interaction between episodic and semantic memory constitutes the constructive process of the ASM. Consequently, manipulating either one's working memory or the process of episodic and semantic memory will influence one's capacity to make moral decisions and lead to moral enhancement.

- How is memory enhancement defined? What does it mean for something to be morally better? Is it possible to answer what morally good is without the agreement on one moral theory?
- What is the relation between one's memory system and moral behavior? Is it possible to enhance moral behavior or sense of morality by means of memory modification?
- Is it moral enhancement morally permissible? Under which condition is it morally permissible?

1.1.8 Special Subjects and Situations

The issues we have considered thus far assume that the subject is an adult with "normal" physical and mental capacities in a "normal" situation, where "normal" refers to statistical normality (see §4.3.1 for different sense of normal). However, we should ask whether there is a different criterion for special subjects or special situations. If there is, what reason is there for such a difference?

Special subjects include children, older people (over 60 years old, according to United Nations, 2001, p. xxviii), patients, animals, and future subjects or generations. First, what leads to a different criterion for special agents is the difference of the moral status they are considered to have. For instance, animals are often regarded as morally inferior to human beings, so the permissibility of animal experiments is controversial, whereas there is general agreement on the prohibition on human experiment. Second, these special subjects are often considered less capable of reasoning and making decisions through independently reflecting on their long-term interests. For instance, the decisions of children are mostly guided by their parents, as they usually lack comprehensive information and are emotionally reliant on their parents' expectations. Do children have authority over themselves? If they do not, who has the authority to decide for them? Take another example in memory intervention: Should we apply therapeutic forgetting therapy on abused children or traumatized elephants in Africa to release them from suffering from traumatic memories?

Special situations include conditions of illness, military service, and medical experiments or research. The context of illness is considered in the distinction between treatment and enhancement and is one of the main issues focused on by this dissertation. What differentiates a situation of illness is the urgency of suffering (see §4.1). In terms of research and experiment to examine the effect of memory intervention, I have discussed the issue of safety in §1.1.2. Should we subject

(special) subjects to tests if this could result in a better understanding of the possibility of memory modification? One issue worth concerning is in order to gain the consent of the subject, how well-informed should the subject be to autonomously participate in the research without influencing the outcome of the experiment. Even if one is well-informed of consent the intervention, does the gain from the research outweigh possible harm to the subjects? The answer may differ from one research to another, but what is the criterion to judge? As for the military context, dampening emotional memory can prevent soldiers from developing PTSD, but can attenuate their moral emotions and affect their moral response. Nevertheless, is normal moral response unnecessary for decision-making in a military context? Does the resulting moral disengagement reduce moral responsibility? It is important to consider long-term effects on subjects under special situations, because these subjects may be constrained by the normative issues that are no longer applicable once they leave the situation (e.g., retired soldiers).

- Is there a different criterion for special subjects (e.g., children, old people, patients, animals, and future subjects) or special situations (e.g., illness, military, and experiment or research)? If there is, what is such a difference based on?
- Do special subjects have a different moral status from "normal subject"? For those who are not capable of expressing their self-interests, who is authorized to decide for them?
- For special situations, what is the moral ground for a situation to be considered morally different?

1.2 Issues at the Social Level

1.2.1 Indirect Coercion

Our conception of something, or our attitude towards it, can be easily affected by other people's conceptions and behavior. If people in one's social environment are taking cognitive enhancers and see this kind of intervention as "normal", it is likely that one will consider such behavior normal and even consider not participating abnormal (see §4.3.1 for more about the interaction between normality and normalization). This kind of social pressure becomes more severe if the CE is accompanied by an advantage in a competitive environment. One may be coerced into accepting CE because of fear of lagging behind. This concerns one's cognitive

liberty (see §1.1.4) in terms of the freedom to take or not to take the cognitive enhancers.

Two kinds of memory manipulation can strengthen competitiveness in one's career. On the one hand, interventions that increase memorizing capacities can bring advantages to those with certain occupations like journalists, lawyers, or students. On the other hand, capacities related to episodic memory, which construct and simulate a particular scenario, support creativity in jobs like writing or making art, which require the ability to recombine elements of memory in a vivid way.

Another form of indirect coercion to memory modification does not necessarily involve memory modification technology. Often we are inclined to believe something if many people around believe it. Likewise, we can be directly or indirectly coerced to remember in a certain way. Such coercion is more common in collective memory, but can also happen in personal memory (e.g., childhood memory). This way of coercion to "remember" certain contents is pervasive and difficult to detect, and triggers the concern of truthfulness (see §1.1.5).

It is also necessary to question why indirect coercion is morally problematic. At a personal level, it affects one's cognitive liberty and autonomy. Furthermore, at a social level, such forces can rapidly alter the collective conception of the intervention. Such processes facilitate a move towards a so-called posthuman future (see also §3.3.4 for evolution of personhood). The idea of transhumanism itself is not morally problematic; however, facilitation through indirect coercion deprives us of opportunities, if there are any, to autonomously decide the direction in which human beings should proceed.

- Why is indirect coercion morally problematic?
- In which way can one be indirectly coerced to take memory enhancers? What kinds of memory enhancement would be easy to distribute under indirect coercion?
- Can direct or indirect coercion to have certain mental content be prevented? How can it be prevented?

1.2.2 The Responsibility to Remember and the Right to Forget

Naturally, we forget and remember. By repetition or by deliberately ignoring certain things, we can train ourselves to increase or decrease our ability to access certain information to a particular degree. If memory manipulation allows us to remember

and forget freely, a further issue arises: Is there a responsibility to remember, or a right to forget?

The "duty to remember problem" (Liao & Sandberg, 2008, pp. 94-95) concerns the idea that there are some memories that are so important that there may be a duty to remember them, for instance, the witnessing of a crime scene, important historical events, shared rules, and so on. However, why is there a duty to remember? Don't we have the right to determine our own minds? We can consider this issue from the perspective of the extended mind thesis (Clark & Chalmers, 1998; Menary, 2010) and the idea of distributed memory (C. B. Harris, Keil, Sutton, Barnier, & McIlwain, 2011; Sutton, Harris, Keil, & Barnier, 2010). The latter considers how the process of memory can involve things outside our skull (e.g., a notebook) or another agent. If it involves another agent, the process of encoding can be either unshared or shared, and the process of retrieval can be carried out in isolation or in collaboration. If there is information that is only accessible to the others through a process involving my brain and cognitive activity, the boundary and ownership of one's mind becomes less clear. If others' minds extend into ours, do we always have the right to forget? When do we have the responsibility to remember? It is noteworthy that the existence of the responsibility is based on the presupposition that the extended process enables retrieval of veridical memory. However, the pervasive of memory distortion over time (after repeated retrieval and reconsolidation) may reduce the responsibility (Lacy & Stark, 2013). If it is the case, research on distinguishing veridical memory from non-veridical one can input to determining one's responsibility of remembering a memory in question.

- Do we have a duty to remember or a right to forget? Why is there such a duty? Don't we have the right to self-determine our own minds?
- When does an obligation emerge? What are the factors determining the responsibility?

1.3 Summary

I have introduced the major normative issues surrounding memory enhancement and modification. As we have seen, these issues are not independent, but entangled: Some are in support of each other, while others are in conflict. In order to determine whether an intervention is permissible under a particular situation, the following are required: (1) conceptual clarification of the issues involved, (2) examination of the criteria and restrictions generated by these issues, and (3) investigation of how the

restrictions between conflicting issues might be resolved. This dissertation deals with the first two parts and focuses on the issue of the distinction between memory treatment and enhancement and the issue of authenticity in the context of memory intervention. The concerns of identity, autonomy, and truthfulness are subsumed by the issue of authenticity.

It is noteworthy that, as I will argue in §4.2.2, the function of the distinction between treatment and enhancement serves as a "moral warning flag", as suggested by Norman Daniels (2000, p. 320). It distinguishes cases in which some normative issues are applicable while others are not. For instance, if an intervention is considered treatment, it requires different consideration in terms of distributive justice. These normative issues are not equally applicable to all subjects in all situations.

As increasing number of researchers have devoted themselves to developing possible candidates for memory enhancers, this chapter examines the normative issues that we need to consider when it comes to the future of our minds. These issues surround the consequences of memory intervention, and mainly concern the alteration of the ASM. ASMs serve as a framework that helps us understand ourselves as embedded in the world: They allow us to look upon ourselves and our lives in the way that we do, and to value and know how we should behave and act.

Chapter 2

Conceptual Tools I:

Memory

2.0 Introduction
2.1 What Is Memory?
2.1.1 The Concept of Memory
2.1.2 The Classifications of Memory and the Memory Process
2.1.3 Alan Baddeley's Model of Working Memory
2.1.4 Endel Tulving's Model of Memory
2.1.5 Autobiographical Memory
2.2 The Representational Theory of Memory
2.2.1 Memory as Ideas
2.2.2 Direct Realism and Representationalism of Memory
2.2.3 Memory as Simulata
2.3 The Constructive Memory
2.3.1 The Constructive Nature of Memory
2.3.2 The Functions of Memory
2.4 Summary

2.0 Introduction

In order to clarify the concept of memory enhancement and to investigate the normative issues of memory interventions, one important conceptual task is to delineate the concept of memory. This chapter aims to provide a theoretical foundation by elucidating what "memory" is. I endorse a view of memory which is based on a representational theory, an adaptive view of constructive memory, and a new perspective on the function of memory.. I will also introduce memory processes and memory systems, which are crucial for the discussions in the following chapters.

The first section focuses on the concept of memory. The term "memory" can denote a variety of things. I will present what "memory" can refer to and define the meaning the concept will carry in this dissertation. Then I introduce the taxonomy of memory systems and the concepts of working memory, episodic and semantic

memory, as well as autobiographical memory (AM). The second section aims to investigate the nature of memory. Some areas of philosophy have long been concerned with the question of how it is that we remember past events. I here present a version of the representationalist account of memory, which allows us to account for our having access to the past event without the presupposition of mental entities. Third, I consider the constructive nature of memory, its relation to future simulation, and the advantages it brings to the organism. This leads to the discussion of the question of memory function. In contrast to the traditional idea of memory function as preservation, a newly proposed function of memory is to provide behavioral flexibility, which promotes the adaptability of the individual.

2.1 What Is Memory?

What does the term "memory" refer to? Although there are some commonalities in the usage of memory languages, e.g., "memory", "memory systems", "remembering", there is no agreement on what these concepts really mean (Dudai, Roediger III, & Tulving, 2007). This lack of consensus and the ongoing debate surrounding this issue result from the different theories of memory (Dudai, 2007; Dudai et al., 2007; Morris, 2007; Moscovitch, 2007; Schacter, 2007). In order to approach the issue of the concept of memory, it is necessary to find commonalities across various forms of memory (Schacter, 2007, p. 23), and to distinguish them from other faculties of human cognition (e.g., perception and imagination). Therefore, this section begins by focusing on different concepts of memory, and then looks into systems, such as working memory, episodic and semantic memory, and autobiographical memory.

2.1.1 The Concept of Memory

"Memory is a big word, encompassing a host of different capacities mediated by functionally distinct components or subsystems that collectively produce the performances we call memory" (N. J. Cohen & Eichenbaum, 1993, p. 16). It is a big term, because it is used to refer to different kinds of memory and because different concepts are involved in the process. The term "memory" often appears along with "learning". It is used in the context in which an system acquires new information, abilities or characteristics, and is able to utilize them. However, the concept remains vauge. On the one hand, the term has been used to refer to different but related concepts, e.g., memory as a cognitive faculty and memory as retrieved information.

On the other hand, there is a long-running philosophical discussion regarding the distinction between memory and other cognitive or physiological states or events. Here, I first deal with the definition of memory: What is included within and excluded from the domain of memory.

What is memory? What features are included within the domain of memory? First, what is not regarded as memory? Fatigue, intoxication, injury, and disease are obviously not. Memory is commonly understood as something acquired through experience and which has the potential to be expressed at another time. However, if we define it as lasting changes in behavior resulting from previous experience, it could also include fatigue, intoxication, injury, or disease, none of which is covered in the concept of memory. What makes memory different from these instances is that it includes changes not only in behavior but also neurocognitive processing. But if we instead define it as the result of short- or long-term changes in neurocognitive processing resulting from experience, some neurological disorders developed through social influences also fall under this definition. Memory, instead, requires a direct link in content between the experience and the changes of mental representation or correlated neurocognitive processing. Accordingly, I define memory as *short- or long-term changes in neurocognitive processing (or mental representations if one adopts a representational account of memory) that directly result from its corresponding experience*.

A certain kind of memory, i.e., episodic memory (which I will introduce later in §2.1.2) shares some characteristics with perception and imagination. The mental states have phenomenal character—there is something it is like for the subject to undergo these experiences. Finding the right distinction between these three has long been the subject of philosophical discussion. Their differentiation will be considered in §2.2, and it is likely that there is no difference between episodic memory and imagination with regard to their nature.

In addition, there are ambiguous cases that are arguably subsumed in the notion of "memory". First, do innate capacities (e.g., the capacity to walk) count as memory? Such capacities, although innate, cannot be acquired without experience (practice). Second, does "external memory", such as Otto's notebook (Clark & Chalmers, 1998), be regarded as memory? In Clark and Chalmers' (1998) well-known thought experiment for exploring the boundaries of the mind, Otto is an Alzheimer's patient who relies on his notebook for dealing with his everyday life. He records newly acquired information in his notebook and consults it for old information. When he decides to go to the Museum of Modern Art (MoMA), he checks the notebook

for information and then goes to 53rd Street, where MoMA is located. Assuming that Otto is still equipped with the capacity to generate neurocognitive processing to integrate the information from the notebook, do these external aids belong to the domain of memory? Some philosophers interested in extended mind or cognition hold that these instances are to be considered part of memory, since they see no difference between resources inside and outside the skull. In their arguments, they make use of different criteria of external memory[4]. Those who object external memory is memory oppose the criteria; e.g., Michaelian (2012) argues that these criteria are not even satisfied by biological memory. I will not go into this issue in this dissertation, but the ethical implications of this issue can be found in §1.2.2.

So far, there is a domain of what kind of processing is involved in memory. However, memory can refer to different aspects of a process. Endel Tulving (2000, p. 36) has listed different meanings of memory:

(1) Memory as *neurocognitive capacity*: This sense of memory refers to the ability to encode, store, and retrieve information, and is used in contexts such as "testing the memory of a subject". Tulving and Craik (2000), for instance, define memory as the ability "to recollect past events and to bring learned facts and ideas back to mind" (p. v). Memory is also defined by Eric Kandel (2007, pp. 9-10) as the ability to acquire and store information, both in daily life and in abstract knowledge.
(2) Memory as *hypothetical store* in which information is held: We refer to this sense of memory in contexts such as "the address is stored somewhere in my memory". The most well-known example is John Locke (2008), who metaphorically describes memory as the "the Store-house of our *Ideas*" (Book II, Chapter X, p. 87).
(3) Memory as the *information stored*: In expressions like "I've lost that piece of memory!" or "memory decay", memory denotes stored information. For example, Morris Moscovitch et al. (2010) regard memory as "a lasting representation that is reflected in thought, experience, or behavior" (p. 305).
(4) Memory as the retrieval of information: This notion of memory is used to denote the result of retrieval or of the "revival of the experience in the past". It is used in sentences such as "I have a vivid memory of the

[4] These criteria include constant access, direct availability, automatic endorsement, and stored information as the consequence of past endorsement (Clark & Chalmers, 1998).

party". Gottfried Vosgerau (Vosgerau, 2010), for instance, argues that only memories that are occurrent (retrieved and activated) have content.

(5) Memory as the phenomenal awareness of remembering something. As we will see later (in §2.2.3), this aspect of memory only exists in episodic memory and Tulving (1983, 1985b) characterizes it as "autonoetic consciousness".

I will not—and I see no need to—argue for a specific use of the term "memory", because these different aspects of memory are consistent with each other and belong to a comprehensive account of memory. Here, I merely want to address the differentiation in order to avoid confusion. Throughout this dissertation, if there is no special emphasis, "memory" is used to refer to *retrieval of information* (the fourth notion listed above). Cognitive capacity (the first notion) will be denoted by the terms "memory capacity" or "memory faculty".

2.1.2 The Classifications of Memory and the Memory Process

Memory is involved in a variety of human phenomena and behaviors: Your short-term memory is engaged when you hand over the right amount of money to the cashier at the supermarket, after he or she has informed you of the amount due; (procedural) memory is what enables you to be seemingly able to play a long-neglected instrument automatically, without having to explicitly think about the procedure; when someone reads the first few lines of a well-known poem, the rest of the text suggests itself to you through your (semantic) memory; and a smell can trigger your longing for family cooking (i.e., episodic memory). This list includes different kinds of memory, which we will look into here.

Memory as a faculty was treated unitary (as noted in Baddeley, 2000b) until William James (1890) distinguished "primary memory", which is the short-lived state in which information reaches consciousness, from "secondary memory", which is regarded as more durable memory concerning what we have learned. This became a starting point for studies on taxonomy of memory. Philosophers and psychologists have been devoted to providing a classification of memory with the hope of reaching a clearer understanding of memory.

Different classifications have been developed in accordance with the interests and focus. Although no one taxonomy of memory is agreed to be correct, these different classifications are mostly compatible. The ways by which types of memory are distinguished from each other depend on the length of time the

information is retained, the degree of awareness of the stored information, and the kind of information that is stored.

Philosophers distinguish memory by content, and come up with a tripartite taxonomy that include experiential, factual, and procedural memory (e.g., Bernecker, 2008; Bernecker, 2010; Sutton, 2010). The content of experiential memory is what we have personally experienced, and it is remembered through experience (e.g., remembering having the 18th birthday party). Thus, this kind of memory involves a first-person perspective as well as qualitative aspects of experience. Autophenomenological reports of factual memory take the form of "S remembers that p", where "S" and "p" stand for the remembering subject and a true proposition respectively (e.g., p can be either "my elementary school teacher took the whole class to a baseball game," or "the capybara is the largest extant rodent on earth"). Unlike the content of experiential memory, "p" is not limited to one's experience. Contrary to these two kinds of memory—which are expressive—procedural memory is not so. Its contents are previously acquired and retained skills, e.g., the ability to ride a bicycle, or to play the piano. When one is practicing the skill (e.g., playing the piano), one recalls the practical memory (of piano playing).

Memory can also be classified by wh-clauses, that is, clauses beginning with "who", "whom", "what", "where", "when", and "why". For instance, we may say "I remember who I met yesterday", "I remember whom I talked to", "I remembered what I ate last night", "I remember where I visited", "I remember when the meeting started", and "I remember why the meeting started late". Bernecker (2008), on the other hand, classified four kinds of remembering based on the objects of the verb "to remember": object- (e.g., remembering the pet dog one used to keep), property- (e.g., remembering the color of its fur), event- (e.g., remembering the dog destroying the trees), and fact-memory (e.g., remembering that the dog destroyed the trees). From my perspective, neither Bernecker's classification nor the categorization according to wh-clauses makes any significant difference to the tripartite taxonomy introduced in the previous paragraph, for these classifications do not describe any distinct kinds of memory that have significant characteristics that are left out by the tripartite taxonomy.

There is another taxonomy of memory that is generally accepted in psychology and neuroscience,[5] although the relationship between the kinds of

[5] There are other taxonomies proposed by psychologists and neuroscientists; see Tulving (1987) for the list.

memory is arguable. This taxonomy of memory is supported by brain research and especially by studies of neural disorders, which act as the evidence of dissociation between different kinds of memory. Thus, I endorse this taxonomy in this dissertation.

First, depending on the length of time for which the information of interest is retained, memory is categorized into sensory, short-term and long-term memory. Sensory memory, which lasts only from milliseconds to seconds, refers to the brief retention of impressions from sensory stimuli. For example, upon hearing the honk of an automobile, the sound seems to be vividly present even though the car has stopped making the noise. The information is stored in the sensory cortex as a short-lived neural trace. Although echoic memory—sensory memory for audition—only retains information for a period of 10 seconds (Sams, Hari, Rif, & Knuutila, 1993), and iconic memory (visual sensory memory) persists for an even shorter while, sensory memory has a capacity of retaining a relatively large amount of information for such a short time.

Short-term memory has a longer time-course and retains from seconds to minutes. Compared to sensory and long-term memory, the capacity of short-term memory is limited. It is usually tested by digit span, that is, the maximum number of items one can hold in memory over a short period of time. George A. Miller (1956) proved that regardless of the content of the items, the number was about seven (the "magical number seven", give or take two). This was later challenged by Nelson Cowan (2001), who proposed that the number should be four, rather than seven.

Information that lasts for a significant time—ranging from days to years, or even a lifetime—is referred to as long-term memory. It is distinguished from short-term memory by two kinds of amnesic syndromes that each shows the dissociation between them: Patients with damage to the temporal lobes and the hippocampi fail in learning and remembering new materials (Milner, 1966) but still have normal function of the short-term memory (a normal digit span), whereas patients with damage to the perisylvian region of the left hemisphere of the brain have a shorter digit span, but a normal long-term memory (Shallice & Warrington, 1970).

Long-term memory is further classified, according to the degree of awareness the subject has of stored information, into declarative and non-declarative memory (shown in Figure 1): The former refers to information that is consciously accessible, whereas the latter refers to information that is not. Requiring no intentional or conscious recollection of previous experience, non-declarative memory is revealed when previous experiences facilitate performance on a specific

task. It encompasses procedural memory (motor and cognitive skills), perceptual priming, conditioned responses between two stimuli, and non-associative learning (habituation and sensitization). Declarative memory is further classified into semantic memory and episodic memory, which will be the topic of §2.1.4.[6]

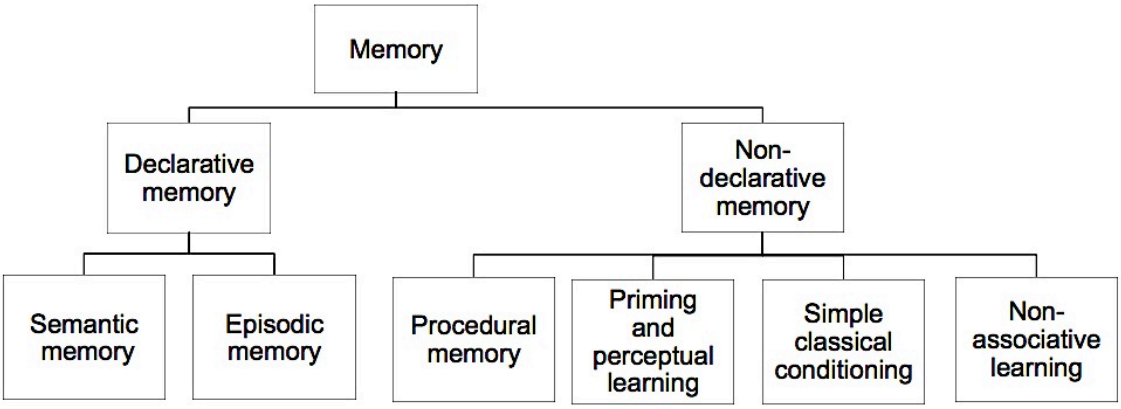

Figure 1. A classification of memory types.
Adapted from "Memory Systems of the Brain: A Brief History and Current Perspective" by L. R. Squire, 2004, *Neurobiology of Learning and Memory*, *82*(3), p. 173.

Before getting into the detail of the classified memory, I provide a brief sketch of the basic processes of memory, including encoding/consolidation, persistence, retrieval, and reconsolidation (see Figure 2). Encoding and consolidation are difficult to separate, both conceptually and empirically. Encoding can be defined as the process by which persistent representations of stimuli or events are formed for later retrieval (Davachi, 2007, p. 138; Hasselmo, 2007, p. 123), whereas consolidation can be defined either as (1) the process that follows perception and enables temporal representations to become long lasting, or as (2) the process at later times, involving a reactivation of correlated neural circuits (e.g., in the hippocampus or neocortical structures) to stabilize representations (Hasselmo, 2007, pp. 125-126). If we adopt the second conception, consolidation may be considered a "post-encoding" stabilization of representations. However, if the first conception is endorsed, "consolidation" conceptually overlaps with "encoding".

[6] As I have mentioned earlier, the tripartite taxonomy, which classifies memory into propositional, experiential, and practical memory, and the taxonomy agreed upon by most psychologists, are compatible. Experiential memory corresponds to episodic memory; factual memory to semantic memory; practical memory to the memory of skill in non-declarative memory.

Whichever of these two conceptions is adopted, distinguishing these two processes is something that requires further empirical research. In Figure 2., only the process of consolidation is shown; encoding can be understood either as the whole process from sensory or bodily input to consolidation, or merely as the process of manipulating information in the working memory.

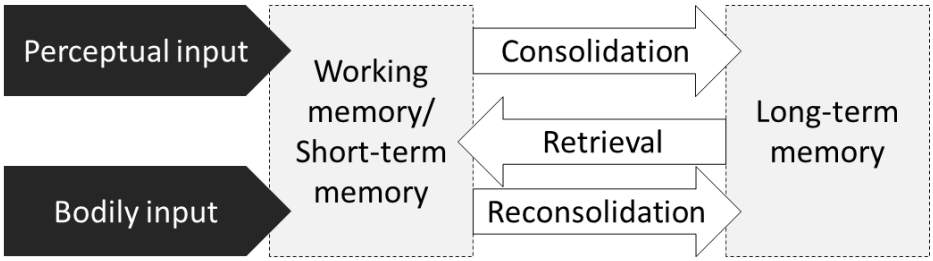

Figure 2. A simplified sketch of memory processing.
Perceptual input includes visual, auditory, olfactory, somatosensory, and gustatory information; bodily input includes interoception, proprioception, and vestibular information.

Second, persistence is the temporal extension of modifications in (neural) representations resulting from experience (Eichenbaum, 2007, p. 193). It replaces the misguided concept of storage that presupposes the Lockean idea of memory as a storehouse (see §2.2.1). As I introduce the constructive nature of memory (see §2.3), we will see why it is misguided. Thus, it is noteworthy that the gray areas in Figure 2 (i.e., working memory/short-term memory and long-term memory) do not represent the storage of memory representation, but two distinct memory process systems.

Third, the concept of retrieval refers to the process involved in utilizing encoded representations, which in some cases (i.e., episodic memory) involves phenomenal features (Roediger, 2000, p. 57). Retrieval can be initiated directly or indirectly by representations associated with target memory. It may involve reconstruction of representations, and the retrieved representations are subject to manipulation. Consequently, retrieval increases vulnerability to modification.

Last, another form of consolidation—reconsolidation follows retrieval (Nader & Einarsson, 2010), which is why the consolidated memory becomes labile after retrieval. Reconsolidation refers to the post-retrieval consolidation, i.e., the process that enables retrieved and (in some cases) modified representations to

stabilize and persist (Alberini, 2005). Two non-mutually exclusive hypotheses explain its function: (1) It allows the new information to be integrated with the old information, and (2) it allows the representation to persist more strongly and for longer (Alberini, 2011).

2.1.3 Alan Baddeley's Model of Working Memory

As memory manipulation could involve the intervention of different kinds of memory, i.e., different memory systems, This section and the next section introduce working, semantic, and episodic memory, which will be mentioned often in the discussions later.

Because working memory has a very close connection to the phenomenal self-model (PSM) and to how the ways in which stored information or mental representations are retrieved—that is, become phenomenal representation—this section, based on Alan Baddeley's model (1992b, 2000a, 2003), introduces working memory and its relationship to short- and long-term memory. Later in §3.2.1, I will illustrate the way in which working memory contributes to one's phenomenal self.

Baddeley (2003) defines working memory as "a limited capacity system, which temporarily maintains and stores information, [and] supports human thought processes by providing an interface between perception, long-term memory and action" (p. 829). According to both the original multicomponent model created by Baddeley and Hitch (1974) and the revised version by Baddeley (2000a), working memory is constituted by four subsystems: the *central executive*, the *episodic buffer*, the *phonological loop*, and the *visuospatial sketchpad*. Their relations to one another are illustrated in Baddeley (2000a, p. 421). The last two components allow memory traces to hold for a few seconds before they fade: Memory traces are, through articulatory rehearsal processes, retrieved and re-articulated. In this way, memory traces are kept active until the information has reached its full capacity. The phonological loop and the visuospatial sketchpad are responsible for speech-based information and visual information respectively.

The central executive, in turn, serves as a control system of limited attentional capacity to bind information from different resources into a coherent episode. There are two components—(1) automatic, habitual control and (2) attentional, supervisory control. The former is the implicit control that is guided by patterns of habit and environmental cues (e.g., driving the regular way back home); the latter refers to attention control and is the "supervisory activating system" (SAS) that intervenes when the former is insufficient. Some, including Baddeley (2003),

have been wary that SAS may just be a homunculus by another name. However, Baddeley believes that by investigating the functions comprising SAS, the homunculus problem can be avoided. As we will see in the next chapter (§3.1), a self-model can substitute the role of SAS without being troubled by the problem.

The last and the most recently proposed component of working memory—the episodic buffer—stores information temporarily and serves as an interface for different information resources, i.e., the visuospatial sketchpad, the phonological loop, and long-term memory. It is attentionally controlled by the central executive and acts similarly to the global workspace suggested by Bernard Baars (1988, p. 42). The addition of the episodic buffer in the model of working memory allows for the information from long-term memory to be "downloaded" to the episodic buffer, and to be manipulated, rather than to be active in long-term memory. Later, the manipulated information can be "uploaded" (encoded or reconsolidated) to long-term memory. The episodic buffer thus provides a working space for the encounter of different representations, and the creation of new representations. As I will show later (in §3.2.1), the integrated information is the content of PSM, and it is accessible to consciousness. Working memory makes a PSM possible by providing such a workspace, holding phenomenal representations temporarily active.

Finally, it is worth noting that working memory is distinct from short-term memory. The content of working memory, unlike that of short-term memory, can originate from sensory memory, long-term memory, or both. The idea of working memory is developed as a limited-capacity storage-space for retaining information over the short term (maintenance) and for performing mental operations on the contents of this store (manipulation) (Baddeley, 1992b). Short-term memory, on the other hand, refers only to the first part of working memory. Short-term memory can be regarded as two of the components of what constitutes working memory—the phonological loop and the visuospatial sketchpad.

2.1.4 Endel Tulving's Model of Memory

Tulving's model (1983, 1985a, 1985b, 2005), distinguishes three memory systems: procedural, semantic, and episodic memory. These are alike in that they all allow for the possibility of utilization of acquired information. However, they deal with different kinds of information, and they have different ways of acquiring and utilizing this information. Procedural memory, allowing the system to learn connections between stimulus and response, handles the "know-how" information, including perceptual-, cognitive-, and motor skills. Semantic memory concerns "the

symbolically representable knowledge that organisms possess about the world" (1985b, p. 2) by making the system create internally representing states of the world. Episodic memory, in order to enable "the remembering of personally experienced events" (1985b, p. 2), deals with the information of "personally experienced events and their temporal relations in subjective time and the ability to mentally 'travel back' in time" (1985a, p. 387).

According to Tulving's model (1985b), these three kinds of memory constitute a bottom-to-top monohierarchy of memory: The systems at the upper level of the hierarchy depend on the systems at the lower levels. The higher levels cannot exist without the lower levels. These three models of memory are companied by three kinds of consciousness: Anoetic (non-knowing) consciousness is associated with procedural memory and results from the organism's capability to sense and to react to external and internal stimulation. Noetic (knowing) consciousness, following semantic memory, enables an introspective awareness of the internal and external world. The object of noetic consciousness is, thus, the organism's knowledge of the world. Last, autonoetic (self-knowing) consciousness correlates with episodic memory and provides the familiar phenomenal flavor of recollective experience characterized by "pastness" and "subjective veridicality". For Tulving, the defining property that distinguishes episodic memory from semantic memory is "mental time travel"—the capacity to mentally project oneself backwards to experienced events (Tulving, 1983, 2005). It allows the subject to, in her mind, "travel back" to an earlier experienced situation, and to mentally re-live or re-experience what has happened.

How are episodic and semantic memories distinguished from each other? Tulving (1972) proposes the following distinction:

> The two systems [episodic and semantic memory] differ from one another in terms of (a) the nature of stored information, (b) autobiographical versus cognitive reference, (c) conditions and consequences of retrieval, and probably also in terms of (d) their vulnerability to interference resulting in transformation and erasure of stored information, and (e) their dependence upon each other. (p. 385)

Table 1 from Tulving (2005, p. 11) offers a detailed list of the common features of two memory systems and the unique feature of episodic memory. I will point to the main differences for Tulving.

Table 1. *The Differences between Episodic and Semantic Memory.*

Episodic Memory	Semantic Memory
Experience-like	Belief-like
Relatively dependent on the context (e.g., time, places, persons, and events) and more self-related	Relatively independent of the context and more world-related (i.e., facts about the world)
Accompanied by a sense of pastness and sense of familiarity	Accompanied by a feeling of knowing
Correlated with hippocampal activities	Correlated with neocortical activities

First, Tulving suggests that episodic memory is distinguished from semantic memory by its function: The function of episodic memory is to enable "mental time travel", while semantic memory does not have such function. The emergence of episodic memory has the main function of increasing behavioral flexibility, as suggested by Suddendorf and Corballis (2007). The functions of different memory systems will be further addressed in §2.3.2.

However, the idea of "mental time travel" is questionable. When I now, in early spring, think about last summer, although I recall the nice walk in the sunshine, I am still quite aware of the cool air of the present. Physically and mentally, I stay within the present circumstance. Remembering differs from the dream state (except in lucid dreaming), in which you lose touch with reality. Besides, when remembering, I still feel that I am present. In fact, it is impossible for someone not to experience nowness. Whatever you experience, you experience it *now* (even in the dreams). Even a confabulator who is convinced that she lives in the 50s still does so *now*.

Second, according to Tulving (2005), episodic memory depends on a remembering "self". The self is what engages in the activity of "mental time travel": "[T]here can be no travel without a traveler" (pp. 14-15). However, what is this "self" that is traveling? Tulving explains that the "self" is what we need for the sake of a complete story of mind to explain phenomenal experience. This issue will be illustrated in the next chapter (§3.1), by drawing on the self-model theory from Thomas Metzinger (2004).

Third, in episodic memory, the present experience is experienced in a way that is linked to a past event. This characteristic allows us to distinguish remembering from perception, imagination, and dreams. It is true that episodic memory is similar in nature to imagination and dreams. Although they are crucially

different, it is not always possible to distinguish these from each other. The distinction between episodic memory and perception can be neatly characterized by the conceptual distinction between presentation, representation, and simulation (detail in §2.2.3). As for episodic memory and dreams, the latter are similar to episodic memory in that they all belong to simulation. However, unlike episodic memory, dreams (with the exception of lucid dreams) are totally detached from the current environment. The most difficult distinction to make is the one between episodic memory and the imagination. Because of their common nature (as simulation) and the constructive nature of episodic memory, I contend that there is no one clear criterion for the distinction between them.

Fourth, unlike semantic memory, episodic memory is orientated in time. Even if one cannot recall the exact time when the event occurred, episodic memory is always accompanied by a sense of "pastness". This is what Bernard Russell (2009, p. 208) adopts in his characterization of memory (see §2.2.1).

Fifth, episodic memory is believed to emerge later in ontogeny and phylogeny. It has often been suggested (e.g., Suddendorf & Corballis, 2007) that episodic memory evolves later than semantic memory, and the question of whether episodic memory exists uniquely in human beings is up for debate (Griffiths, Dickinson, & Clayton, 1999; Suddendorf & Corballis, 2007). Moreover, compared with the susceptibility, in episodic memory, to forgetting and modifying, semantic memory is less vulnerable to change.

Last, there is something else worth noting: It is generally believed that semantic memory is independent from the spatial and temporal contexts in which it is acquired, while episodic memory is considered to relate to a specific time and space. For instance, the recipe of a lemon cake is held in semantic memory, while the memory of being in the kitchen, making the lemon cake for a friend birthday is part of episodic memory. Moreover, it is commonly held that episodic memory handles personal information, while semantic memory concerns the world. That is, episodic memory, unlike semantic memory, is context-dependent. However, contextually dependent personal information also resides in semantic memory. It is then called "personal semantic memory" (Kopelman, Wilson, & Baddeley, 1989). This concept is defined as "factual knowledge about a person's own past" (e.g., addresses where lived, names of teachers/friends/colleagues at work, etc.) (p. 726). I summarize the differences between episodic and semantic memory in Table 1.

Tulving (1993, 1995) proposes the *serial-parallel-independent model* (SPI) depicting the functional relations between the processing of the two memory

systems. First, according to the SPI model, encoding proceeds in a serial form. First, the information is encoded in semantic memory, before reaching the encoding process of episodic memory. Then, the information persists (or is stored) in parallel mode: The information from one memory system can be lost without the loss of the corresponding one in the other memory system. Last, they are retrieved separately. Incorporating the SPI model enriches the diagram shown in Figure 2 (i.e., the gray box labeled long-term memory).

2.1.5 Autobiographical Memory

As we have seen in the last section, episodic and semantic memory cannot be neatly distinguished from one another. Instead, these two memory systems are interactive and interdependent. Patients who suffer from deterioration of one of the memory systems provide subjects for investigating the way in which one functions without the other. However, the everyday memory of healthy subjects under most circumstances, instead of laboratory settings, is the result of the interaction between episodic and semantic memory. If one looks into one's everyday memory, one will find the characteristics of both episodic and semantic memory. For instance, when one recalls the trip to Weimar, such a recollection is constituted by experience-like memory (e.g., the cream-yellow courtyard of the Goethe Residence) as well as by belief- or knowledge-like memory (e.g., the knowledge that Weimar is located in central Germany or that Johann Wolfgang von Goethe was the author of *The Sorrows of Young Werther* and was buried in Weimar).

In addition, episodic and semantic memory can't be neatly differentiated by the characters of the information retrieved. Traditionally, the contents of recollection of episodic memory are considered to be self-related, while the contents of semantic recollection are world-related. The former are recollections of particular personal events; the latter hold facts about the world. However, as I mentioned earlier, personal semantic memory is an exception to this division. Resulting from semantic memory, it concerns the general facts of oneself.

Therefore, an increasing number of empirical studies have investigated the interdependence of episodic and semantic memory. First, according to Tulving's SPI model of memory processing, encoding proceeds in a serial form: Information goes through semantic processing before reaching episodic memory. Consequently, impaired semantic memory results in a failure of the encoding of episodic memory (but not vice versa). Moreover, Tulving (1983) has noticed the interdependence between the two, which he considers to be variable: The interdependence is

pronounced in some cases and negligible in others (p. 26). For instance, the generalization of memory (Lacy & Stark, 2013, pp. 653-654)—the retrieved memory becomes more semantic and less episodic with the passage of time—may result from the different interaction between episodic and semantic memory at retrieval and reconsolidation. In addition, Greenberg and Verfaellie (2010), when investigating the relation between the two memory systems at encoding and retrieval, conclude that episodic memory facilitates the encoding and retrieval of semantic memory, and the latter can also facilitate the encoding of the former.

Autobiographical memory (AM) is one of the evident examples of the interplay of the two memory systems. There are distinct definitions of AM: Baddeley (1992a) defines it as "the capacity of people to recollect their lives" (p. 26); Rubin (1986, 1999), however, contends that the definition should not be set a priori; Williams, Conway, and Cohen (2008) defined AMs as "episodes recollected from an individual's life" (p. 22). Here, rather than, as Baddeley does, regarding AM as a kind of neurocognitive capacity I will side with the last definition and treat it as a kind of recollection (see different conceptions of memory in §2.1.1). I define AM as *a recollection comprised by self-related information*. According to such a definition, AM is fundamentally different from episodic and semantic memory: The former refers to the retrieval of information; the latter consists of information-processing systems.

In other words, AM is the result of the interplay of episodic and semantic memory. As mentioned earlier, episodic memory is considered the memory system responsible for personal information, and it thus contributes to AM. However, it has been shown that semantic memory also plays an important role also in AM. Semantic memory provides AM not only with personal semantic content, but also with schema for integrating episodic details (Irish & Piguet, 2013). This view is supported by the study of patients who suffer from semantic dementia (SD), which is a variant of fronto-temporal dementia. It is characterized by a range of symptoms, including a deficit in semantic memory. Maguire and colleagues (2010), studying AM in SD patients, have found that autobiographical impairment deteriorates as semantic impairment worsens.

Later, in §3.2.4, I will also introduce the concept of the autobiographical self-model (ASM), to which AM is closely related. In this dissertation, because I am mainly interested in how memory, as a kind of knowledge, shapes and contributes to our life and self-knowledge, I will only focus on declarative memory, and leave out procedural memory. I will also deal with short-term memory, but only as a

component of working memory, which is integrated into the theory of self-model, presented in the next chapter (§3.2.1). As you will see later, the emphasis will lie on AM and how episodic and semantic memories mutually contribute to it. AM along with the autobiographical self-model, will function as important conceptual tools throughout the remaining chapters.

2.2 The Representational Theory of Memory

After differentiating different concepts of memory and classifying different memory systems, I consider the nature of memory in this section. Starting with the traditional philosophical theories of memory, the debate between direct realism and representationalism of memory is reviewed. Then, based on a representational theory of memory, memory is characterized by *(self-)simulata*.

2.2.1 Memory as Ideas

According to his *Essay Concerning Human Understanding*, first published in 1690, John Locke's (2008) idea of memory is considered to be a version of representationalism. According to Locke, memory is "the store-house of our *ideas*" (p. 87, Book II, Chapter X), where ideas refer to "whatsoever is the object of the understanding when a man thinks" (p. 16). Memory is regarded as one way of retention, namely "the power to revive again in our minds those *ideas*, which after imprinting have disappeared, or have been as it were laid aside out of sight" (p. 87). Memory serves as a "repository" (p. 87) from which to supplement the narrow mind of man: Because we cannot consider too many ideas at once, memory allows us to store them for future use. For Locke, memory is regarded as revived perception, although when revived, they are painted anew: "some with more, some with less difficulty; some more lively, and others more obscurely" (p. 87). The idea of memory viewed as a copy or representation of perception begins from this account by Locke.

David Hume, in his *Treatise of Human Nature*, first published in 1739, further distinguishes "representations" (i.e., "ideas", according to Locke) into "impressions" and "ideas" (Hume, 2004, p. 10). For him, the former refers to "sensations, passions and emotions", while the latter refers to "the faint images of these in thinking and reasoning". The two can be easily differentiated through the perception of the difference between feeling and thinking. According to Hume, every impression is accompanied by a corresponding idea, and every idea has a

correspondent impression. As an empiricist, he holds that all ideas derive from reflections, with corresponding impressions, which they represent. They are different not in nature, but only in degree. For instance, the impression of red that strikes our eyes, and the idea of red that we form in the dark, differ in degree of *force* and *vivacity*.

However, in some circumstances (e.g., in a fever, or in madness) impressions and ideas may approach each other, and the distinction becomes less clear. Memory is such a case. According to Hume (2004, p. 14), following the presence of an impression, the appearance of an idea either "retains a considerable degree of its first vivacity" or entirely loses its vivacity. Memory as the former high-vivacity idea is thus distinguished from the imagination, the latter being the "perfect idea". He metaphorically suggests that memory, with the preservation of the vivacity of impression, presents more detail than the imagination:

> It is evident at first sight, that the ideas of the memory are much more lively and strong than those of the imagination, and that the former faculty paints its objects in more distinct colours, than any which are employed by the latter. (Hume, 2004, p. 14)

Memory is therefore illustrated as residing somewhere between impressions and (perfect) ideas. There is another way to distinguish memory from the imagination. Hume writes "the imagination is not restrained to the same order and form with the original impressions; while the memory is in a manner tied down in that respect, without any power of variation" (2004, p. 14). Memory preserves the original form of the object presented, and its main purpose is to preserve not the simple ideas but their order and position. This distinction is not tenable, as I will show in the next section that under most circumstances memory neither preserves its form nor serves such purpose.

Bernard Russell, in his 1921 work *The Analysis of Mind*, follows the representational idea espoused by Locke and Hume. Diverging from his previously direct realist account, he claims that memory demands an image (2009, p. 237), where images are regarded as "copies" of sensations or experiences in the past (p. 101). In contrast to Hume's distinction of memory and the imagination, he claims that because of the existence of the cases in which we distrust vivid images (e.g., due to the influence of fatigue), we can know the past in a way that is independent from images. This is supported by the "feeling of familiarity" and the "sense of

pastness" that accompanies the image (p. 208). The former, capable of degrees, informs us of the accuracy of the image and leads us to trust it; the latter assigns its place in the time-order. These two feelings distinguish real memory images from imagined ones.

Thomas Reid (1878) rejects the use of the concept of ideas in this context, and advocates a direct theory of memory. One of Reid's objections to Locke and Hume is the difficulty of distinguishing the work of memory from that of the imagination. Hume's differentiation of memory and imagination relies on the degree of "force and liveliness". Such a distinction is not sufficient for Reid: He compares the vivacity of memory and imagination in one's striking one's head against the wall, and one's touching one's head against the wall. If Hume is right, the latter must be memory (p. 227). As for Russell, whom Reid is incapable of criticizing, *déjà vu* provides a counterexample for the differentiation between memory and imagination. *Déjà vu* refers to the phenomenon of having the feeling of already having experienced something that actually now occurs for the first time. In such cases, feeling of familiarity and pastness occurs without the work of memory. More memory distortions (*§2.3.1*) can be considered counterexamples here.

As a direct realist, Reid refutes the existence of ideas, images, and representations. In *An Inquiry Into the Human Mind on the Principles of Common Sense* (1769) he provides the following illustration:

> Philosophers indeed tell me, that the immediate object of my memory and imagination in this case, is not the past sensation, but an idea of it, an image, phantasm, or species of the odour I smelled: that this idea now exists in my mind, or in my sensorium; and the mind contemplating this present idea, finds it a representation of what is past, or of what may exist; and accordingly calls it memory, or imagination. This is the doctrine of the ideal philosophy […]. [M]emory appears to me to have things that are past, and not present ideas for its object. [W]hen I remember the smell of the tuberose, that very sensation which I had yesterday, and which has now no more any existence, is the immediate object of my memory. (1769, pp. 32-33)

On the one hand, in Reid's view, if one is committed to the existence of ideas, ideas cannot be constantly present, because if ideas were treated as direct causes, then we would constantly perceive. If Locke's storehouse of ideas is regarded as a metaphor, Locke is required to explain how it is that identical ideas can reappear, for he claims

that keeping identity over time requires continuous existence (Copenhaver, 2009). On the other hand, according to Reid, the object of memory is something in the past, rather than in the present, for if its object is something in the present, it will be perception rather than memory.

Furthermore, in *Essays on the Intellectual Powers of Man*, Reid endorses the idea, which is later coined by Sydney Shoemaker (1970, p. 269) as "previous awareness condition", that in order to remember a past event, it is necessary that the same individual has observed or experienced that event. He then questions how it can be that one can reach certainty, based only on images and ideas, regarding their veridicality. He illustrates this point as follows:

> [A]ccording to that theory, the immediate object of memory, as well as of every other operation of the understanding, is an idea present in the mind. And, from the present existence of this idea of memory I am left to infer, by reasoning, that six months or six years ago, there did exist an object similar to this idea.
>
> But what is there in the idea that can lead me to this conclusion? What does it bear of the date of its archetype? Or what evidence have I that it had an archetype, and that it is not the first of its kind? (Reid, 1786, p. 274)

Therefore, he believes that to posit the roles of images or ideas is problematic for a theory of memory. But, according to Reid, how are we able to tell memory from perception and imagination?

On his account, memory is immediate in the same way that perception is immediate. Memory is always accompanied by "the belief of that which we remember", just like perception is accompanied by "the belief of that which we perceive" (1878, p. 212), and any person with a sound state of mind cannot confuse memory with imagination. For Reid, to prevent circularity, memory is unaccountable (or according to the interpretation of Senor (2009), unanalyzable) in the sense that no reason can be given to explain why I have this knowledge and not another.

However, as we will see later on, memory distortion is very common even in those with a sound state of mind. Under some circumstances, the belief that is supposed to accompany memory, according to Reid, can accompany imagination. In §2.3, I will focus on the nature of memory and show how memory distortion can be central and beneficial to human cognition.

2.2.2 Direct Realism and Representationalism of Memory

Reid's account of memory differs from those of Locke, Hume, and Russell in that Reid rejects the existence of ideas or representations. According to him, when we remember, we have direct access to past events (e.g., to the smell of the tuberose encountered yesterday). This difference highlights the debate between direct realist and representationalist accounts of memory. In this section, I will present this debate and conclude with a version of the representationalist take on memory.

Let's begin with an example. When I think of the shooting stars, I saw a few years ago, the light, the color, and the shape spring to mind. How is my present act of remembering—related to the past experience—able to "revive" things that do not exist anymore (e.g., the shooting star that lasted less than a second)? Direct realist and representationalist accounts of memory hold two distinct views on the nature of this relation.

The debate between direct realism and representationalism of memory starts from the dispute over whether we have direct or indirect access to past events (Sutton, 1998, p. 280). According to the former, in the act of remembering, we are in direct contact with past events. By contrast, representationalists claim that when one recalls a past event, what one is aware of is an idea, image, or representation. The dispute also springs up regarding whether memory necessarily requires the mediation of a memory representation (Lawlor, 2009): Direct realists contend that our access to the past is immediate, while representationalists claim that memory requires the mediation of memory representations. Direct realism and representationalism each have their own obstacles. I will consider the major objections against them and argue for a version of representationalism.

The classic view of representationalism (as cited in Bernecker, 2008, p. 65) endorses the following distinct claims: (1) Memory representations are mental entities which serve as objects of memory; (2) a memory representation of something X (e.g., the taste of a piece of chocolate) shares numerous properties with some prior perception of X (e.g., the sweetness and bitterness); (3) a memory representation of something X is causally linked with some prior perception of X (e.g., the memory representation of the taste of chocolate is in a way causally linked with the past perception of that chocolate). This version of representationalism of memory is problematic. First, it results in a form of dualism: Representations that are, ontologically, mental entities sharing properties with prior perceptions are incompatible with naturalistic ontology. However, as we will see later, a version of

representationalism, which considers representation as neural processing, avoid such problems.

Second, there is the worry that a representationalist stance on memory leads to epistemological skepticism. According to this theory, we cannot be directly aware of what we experienced in the past, but can only have access to the past through the mediation of the present internal representations of past things. Since what we can know about the past is restricted to representations, the question is how we are to know that our memory experiences faithfully portray the past? This characteristic of the representationalist account, which is seen by critics to be a defect, nevertheless describes the character of human memory. Our understanding of the past truly is restricted. Without any confirmation from reliably recorded information (e.g., diaries, photographs, or home videos), left-behind traces (e.g., plan ticket, receipt, beer bottles), and memory episodes shared with others, there is no way that we can confirm the veracity of our memories.

Direct realist accounts of memory, however, are motivated by the phenomenology of remembering: When we remember something, what we are aware of is merely that thing, and nothing more (§2.2.1; Bernecker, 2008, p. 67). For instance, when recalling the food you tasted last night, what you are aware of is the taste and other sensation you experienced, instead of "representations". Proponents of direct realism of memory include Reid (1769, 1786, 1878) and Sven Bernecker (2008, 2010). One of the main problems of direct realism is to accommodate an explanation of how it is that we have direct access to past experiences that no longer exist. As mentioned above, in §2.2.1, Reid holds that we have immediate knowledge of the past by virtue of an act of the mind. However, he does not explain what kind of act this is and in what way this gives us immediate knowledge of the past. He seems to attribute it to the unaccountable or the unanalyzable (as cited in Senor, 2009).

Second, as direct realist accounts of perception face the argument from illusion, so direct realist accounts of memory are required to respond to the fact of memory distortions: If one perceives something with a certain property in virtue of one's direct and immediate awareness of its property, how is illusion or hallucination even possible? An account of perception must, in order to account for the possibility of error (e.g., illusion and hallucination), allow for the possibility of misrepresentation. Likewise, memory is not infallible: We often make mistakes in remembering. For instance, a national study on how accurately Americans (3,000 people across 7 U.S. cities) were able to recall their memories of the 9/11 attacks

shows that three years after the attack, people's recollections of the details of the events were only around 50% accurate (Hirst et al., 2009). The direct realist account of memory, which proposes that memory provides us with immediate knowledge of the past, does not allow for the possibility of memory going wrong, while the representationalist account does.

However, Bernecker (2008, 2010), classifying himself as a direct realist, recognizes the existence of representation in remembering.

> Though remembering something may require the having of memory-data, there is no reason to suppose we are aware of these memory-data themselves. I am aware of a past event by internally representing the event, not by being aware of the internal representation of the event. Memory-data do not function as the primary objects of awareness, but are merely the vehicles of the remembered information. (2008, p. 75)

For Bernecker, it is acceptable to say that memory is indirect in the sense that it involves a series of causal intermediaries between past events and present memory experiences, but not in the sense that the object of awareness is something other than the experience in the past. That is, Bernecker (2008) places the distinction between representationalism and direct realism in that which we are directly aware of, rather than in the existence of representations.

If we adopt this distinction, some versions of representationalism and direct realism are compatible with each other. As critics of representationalism have noted, there is no two-step procedure in which we first become directly aware of a representation and then, indirectly, infer this to past experience (Sutton, 1998, 2010). In addition, modern representationalists reject the existence of mental entities as representations (Seager & Bourget, 2007). Representations are treated as physical entities or seen in terms of neural processing.

The concept of representation can be analyzed in terms of a tripartite relationship between *representation*, *representandum*, and *representatum* (Metzinger, 2004, pp. 20-21). The representandum refers to the object of representation, which can be external facts (e.g., source of food) or internal facts (e.g., blood sugar level). The representatum is the internal state that carries information about the object. The representation is the process by which the system generates the representatum to create an internal depiction of the representandum. Since representation is the information or neural processing ("vehicle"), it does not

itself reach one's awareness. What one is directly aware of is, rather, the content of the representation. If one adopts this version of representationalism, the conflict between representationalism and direct realism appears insignificant, for according to this account, we can have direct awareness of the past by virtue of a representation in the present (Sutton, 2010).

As we will see in the next section, memory, unlike perception, involves a special kind of representation, that is, simulation. The object of simulation is a counterfactual external or internal state, instead of the current state. To apply the tripartite relationship on memory, the simulandum is the content of a past representation (e.g., what I experienced on my last trip); the simulatum refers to the internal state (e.g., what I can potentially experience at retrieval); the simulation is the process that the system generates in order to create the internal depiction of a past event.

2.2.3 Memory as Simulata

If we endorse the representationalist theory and grant that memory exists as a kind of representation, when we think of an event in the past, what we are actually aware of is the content of representation, just like in perception. However, how is memory to be differentiated from perception? How can a representationalist theory account for the difference between perception and memory? In this section, I will first illustrate the phenomenal differences between the two, and then introduce Metzinger's concepts of representation and simulation. The phenomenal differences can be characterized by the differences between representation and simulation.

When we think of the family dinner last week, or of the face of a close friend, what we experience shares some resemblance to the experience we had before: the smell of the food, the taste of the wine, and the sound of people talking seem to be "revived" in the mind. Terms such as "re-experience" and "mental time travel" are employed to illustrate recollection. Nevertheless, these descriptions, emphasizing the similarity between memory and the experience it represents, ignore the significant differences between them.

Hume (2004) describes the differences between impressions and ideas in terms of "force and liveliness" (p. 10). This description might be able to shed light on the issue in question. Impressions characterized by "force and liveliness" are involved in the experience of something, e.g., hearing, smelling, or seeing. By contrast, ideas without "force and liveliness" are related to thinking as opposed to experiencing. This differentiation of impressions and ideas characterizes the

difference between perception and imagination. After we have had an impression, the impression may "make its appearance as an idea" (p. 14) and retain parts of its initial vivacity and forcefulness. According to Hume (2004), a memory image is "somewhat intermediate betwixt an impression and an idea" (p. 14). It is, by definition, fainter than impressions, but more vivacious and forceful than images of the imagination. Our imaginings do not become percepts just because they are forceful and lively enough to fool us. However, Hume's account seems to suggest that they should. Yet, even if we occasionally do make that mistake, it remains a mistake.

Among the differences between percepts and images considered by Colin McGinn (2004), we find the property of "saturation". According to his description, "the percept represents the world as dense, filled, continuous; but the image is gappy, coarse, discrete" (p. 25). He emphasized the difference in phenomenology that there is a manifested quality that can be found at every point of the phenomenal visual field in experience, and that cannot be thus found in the image. The example McGinn provides is that in forming an image of your mother you will select certain features that are sufficient to make it an image of your mother. By contrast, when you see your mother, there are no blank spaces. It is thus that percepts are "saturated" and have the quality of "phenomenal plenitude" while images are "unsaturated" and contain a quality of "poverty" (p. 26). Regardless of McGinn's attempt to draw a sharp distinction between percepts and images, his concept of "saturation" successfully describes the phenomenal difference between memory and perception.

Metzinger's (2004) concept of (self-)simulation, as distinguished from (self-)representation, and the functional constraints of "offline activation" and "transparency" characterize the differences between memory and perception. In the rest of this section, I will consider (1) the distinction between *simulation and representation*, (2) the distinction between simulation/representation and *self*-simulation/*self*-representation, and (3) the distinction between *mental and phenomenal* simulation/representation. After introducing the constraints of offline-activation and transparency, I will show how the concepts of mental/phenomenal simulation/self-simulation can be employed to explain memory.

A mental representation of a (bio-)system is the information processing that generates an internal state (representatum) that, in turn, depicts an aspect of the current state of the external world (Metzinger, 2004, p. 21). In contrast to sensory-driven content of representations, simulations—which are "'virtual' representational

processes"—generate "possible phenomenal worlds" (p. 43). A simulation is an information process that generates a simulatum, which simulates a counterfactual situation of the external world for the (bio)system. Examples of such states include mind wandering, dreams, hallucinations, memory, and future thinking.

Unlike mental representations, which rely on sensory input and whose content is stimulus-related, mental simulation, triggered by external or internal stimuli, does not involve the kind of activation of sensory correlates involved in mental representation. Accordingly, simulata are not stimulus-related (p. 44). This kind of representation allows the (bio-)system to have "a larger inner behavioral repertoire" (p. 49). Mental simulation allows an organism to process abstract information and to carry out internal simulations of complex, counterfactual sequences of events such as planning and memory, which results in a more complicated behavioral pattern and in the capacity to act (see more in §2.3).

Next, the concepts of mental self-representation and self-simulation are in contrast to the concepts of representation and simulation (Metzinger, 2004): Whereas the object of the latter is the external world or the current or counterfactual state of the external world, the object of the former is the system as a whole, or the current or counterfactual state of the system. That is, a mental self-representatum is an internal "image" of itself constructed by the system (pp. 265-267). Likewise, mental self-simulata, as *possible selves* (p. 280), are utilized by (bio-)systems in order to process abstract information about themselves so as to achieve their (biological) goals (pp. 279-282). Mental self-simulata are activated independently of actual internal input, and they therefore do not have a high covariance with actual inner states. They are embedded in the "phenomenal possible worlds" generated by simulations.

Last, mental and phenomenal representations are distinct from one another: The former are "the states possessing the dispositional property of *becoming* globally available for attention, cognition, and action control in the window of presence defined by the system" (p. 42), and the latter are the states that are *currently* globally available. That is, a phenomenal representation can become the object, or representandum, of another, sub-symbolic or symbolic, high-order representational process. Likewise, a mental (self-)simulation becomes globally available when it becomes phenomenal. For a mental representation to become a phenomenal representation, one functional constraint that is essential to satisfy is *transparency* (see §3.1.4 for more details on functional constraints for phenomenal representation/self-model).

Phenomenal transparency is a necessary functional constraint for a mental representation/simulation to become phenomenal. A representation becomes transparent if its earlier processing stages are unavailable for attentional introspection; that is, if it becomes unavailable to the sub-symbolic meta-representations operating on currently active inner models (Metzinger, 2004, p. 165). It is worth noting that transparency is a phenomenological property of conscious representation, and being inversely correlated to the degree of attentional availability of earlier processing stages, it can be more or less transparent. This degree of transparency and opacity is what Hume (2004) characterizes as "force and liveliness", and what McGinn (2004) describes as saturated or unsaturated.

As to the constraint of offline activation, this refers to the way in which phenomenal (self-)simulations are generated, and to current external or internal states. It is due to this constraint that not only present stimulus-related objects, but also past and future episodes, can be generated. Accordingly, the "*historicity of one's own person*" is here both cognitively and phenomenally available (Metzinger, 2004, p. 180). It is noteworthy that simulata, which satisfy the constraint of offline activation, are typically opaque (p. 179).

How are these conceptual tools related to our present concern—memory? First, the differences between perception and memory are characterized by the distinction between representata and simulata. Perception realized by representation corresponds to the current states of the world, while simulation allows the organism to generate an inner state of the past and allows for the emergence of memory. Besides, the difference between the phenomenal quality of perception and memory is characterized by the phenomenological concept of phenomenal transparency. The vividness of the current perception you have is a result of representation, which involves the activation of sensory correlates stimulated by actual and external stimuli. On the other hand, recall your last vacation: The opaque quality is a result of the lack of full-blown correlated activation. Consequently, the degree of transparency is expected to increase with the increase of external cues or potential brain stimulations.

Second, simulation and self-simulation are distinguished by the objects of simulation processing—the counterfactual state of the external world or the system. This distinction also enriches the distinction of semantic and episodic memory. For instance, when I recall my last attendance at a conference, I can retrieve the semantic memory of what happened during the conference, as well as the episodic memory of the experience I had while attending it. The former results from

simulation, while the simulatum is the counterfactual state of the world. The latter is considered self-simulatum, for it is the counterfactual state of myself.

Last, as will be shown in §3.1, a mental (self-)simulation turns into a phenomenal (self-)simulation as it is integrated into the current active phenomenal self-model. That is, we can consciously and cognitively access declarative memory only when it appears as a part of the phenomenal self-model. Details on the relation between (self-)simulation and the phenomenal self-model will be discussed in §3.2, in which I develop the concept of (phenomenal) ASM.

2.3 Constructive Memory

We have examined the concept of memory and seen how it can be understood as (self-)simulata in a representatonalist theory. This section investigates the mechanism of the formation of a memory—how recollection occurs—and the function of memory—what determines the function and malfunction of memory. These issues surround the constructive nature of memory.

2.3.1 The Constructive Nature of Memory

In *How the Mind Forgets and Remembers: The Seven Sins of Memory*, Daniel Schacter (2001) introduces seven shortcomings of memory. The first three sins—*transience*, *absent-mindedness*, and *blocking*—refer to different kinds of forgetting. They refer, respectively, to the gradual forgetting over time, forgetting due to insufficient attention, and temporary inaccessibility. The following three sins—*misattribution*, *suggestibility*, and *bias*—are distortions. They are the attribution to an incorrect time, place, or person; the tendency to incorporate invalid information (e.g., misleading information suggested by others), and the vulnerability of recollecting the influences of present knowledge, beliefs, and feelings. The last one—*persistence*—refers to the involuntary recollection of a fact or event that one would prefer to forget. Elizabeth F. Loftus and colleagues study cases and different forms of memory distortions. They have, for instance, investigated memory distortions in eyewitnesses (Wells & Loftus, 2003), created childhood memory (Loftus, 1997), errors in AM (Hyman Jr & Loftus, 1998), and false memories of political events (Frenda, Knowles, Saletan, & Loftus, 2012).

The sins of memory have evoked the question of the nature of the memory systems. Are our memory systems fundamentally defective, or do systems as such serve other functions? Schacter (2001) argues for the latter. He argues that these

sins are regarded as the by-products of adaptive features of memory (p. 184). The studies of sins have led to the progressively popular view of memory that is referred to as the constructive nature of memory.

There are three aspects of memory that overturn the traditional view—or folk psychology—of memory. According to the traditional view, memory is regarded as a cognitive function and serves to represent past events. First, memory is considered a matter of the past. Its function lies in enabling the past and vanished events to be mentally revived at a later time. Second, memory is seen as the storehouse where we keep representations and to which we withdraw during recollection. Third, a well-functioning memory system is considered one that can faithfully reproduce past facts or events. According to such a view, the "sins" that Schacter introduces are considered defects of a memory system. However, an alternative view has been proposed against its predecessor, and empirical studies have accumulated in support of it. For the remainder of this section I will focus on the adaptive constructive nature of memory and future episodic simulation, and the function of memory will be discussed later (in §2.3.2).

Frederic C. Bartlett, in his classical work *Remembering*, first published in 1932, suggests that we get rid of the idea that memory is reduplicative or reproductive (1995, p. 204). Instead, he contends that memory

> […] is an imaginative reconstruction, or construction, built out of the relation of our attitude towards a whole active mass of organized past reactions or experience, and to a little outstanding detail which commonly appears in image or in language form. (p. 213)

And such construction relies on schema, where he defines a schema as "an active organisation of past reactions, or of past experiences" (p. 201).

Chris Westbury and Daniel Dennett (2000) also address such constructive ideas of memory. They argue that "[w]hat we recall is not what we actually experienced, but rather a reconstruction of what we experienced which is consistent with our current goals and our knowledge of the world" (p. 19). According to them, recollection is nothing more than a plausible story we come up with within our biological and historical constraints.

Berstein and Loftus (2009), while aiming to distinguish true memory from false memory, make this remark:

> In essence, all memory is false to some degree. Memory is inherently a reconstructive process, whereby we piece together the past to form a coherent narrative that becomes our autobiography. In the process of reconstructing the past, we color and shape our life's experiences based on what we know about the world. (p. 373)

For more than a decade, cognitive neuroscience on memory distortions has provided evidence in support of the constructive view of memory.

Schacter and colleagues (1998) propose a "constructive memory framework", according to which representations of new experiences are conceptualized as "patterns of features, in which different features representing different facets of the experience" (p. 290). Such representations (i.e., simulations) are distributed across different regions of the brain. Accordingly, retrieval is realized by a process of pattern completion, which allows a subset of features to comprise a past experience. That is, memory representations are not stored intact in a "Store-house" and drawn out on retrieval, as Locke (2008) claimed, but they are stored as elements that are then reconstructed. The construction of memory is constrained by different factors, including the present goal of the organism (Conway, 2005), the current environment (Anderson & Schooler, 1991), and the prior knowledge (Hemmer & Steyvers, 2009).

2.3.2 The Functions of Memory

Last, I introduce the idea of the constructive nature of memory. Current developments have moved into the direction of addressing the question whether such a constructive nature of memory implies a fundamental flaw of the memory system. Recent studies have focused on the link between the constructive nature of memory and future thinking.[7] If the constructive nature of memory is in the service of future thinking in order to achieve future goals, resulting distortions are to be considered the by-product of the function of memory.

Traditionally, we consider memory to be something that only refers to the past. However, what matters to an organism are the *current* and *potentially future*

[7] Some have used the term "prospective memory" to refer to the (self-)simulata with the future aspect. However, to prevent the confusion resulted from the concept of "memory" which involves the (either true or false) belief that the content of memory refers to something in the past, I use the term "future thinking". And after I have introduced the concept of autobiographical self-model (ASM) in the next chapter, it is illustrated as part of an ASM.

states of the environment and the organism. Then, in which way does memory contribute to an organism? As Westbury and Dennett (2000) illustrate:

> The whole point of brains, of nervous systems and sense organs, is to produce future, to permit organisms to develop, in real time, anticipations of what is likely to happen next, the better to deal with it. The only way—the only non-magical way—organisms can do this is by prospecting and then mining the present for the precious ore of historical facts, the raw materials that are then refined into anticipations of the future. (p. 12)

For them, memory must be based in utility: Only by retrieving information that increases the likelihood of achieving adaptive goals will organisms gain advantages from memory (p. 13). Pascal Boyer (2009) addresses the same point:

> Obviously, we have memory because of evolution, because of the kinds of organisms we are, as a consequence of our evolutionary history. Now the past does not affect an organism, except through its consequences for present circumstances. So if we consider memory as a biological function, we are led to consider that memory is certainly not about the past but about present and future behavior. Memory has a biological function to the extent that it serves to organize current behavior. (p. 3)

Therefore, memory not merely concerns the past, but it also contributes to present and future ends. But, how does memory achieve that?

Schacter and Addis (2007a, p. 778; 2007b) put forth a *constructive episodic simulation hypothesis*, according to which the constructive nature of episodic memory is partially attributable to the role of allowing us to mentally simulate our personal futures. That is, memory is adaptive in allowing us to employ past experiences to in such a way as to enable simulations of possible future episodes. Suddendorf and Busby (2003) address the same idea, when answering why human organisms have evolved a sometimes unreliable constructive system. They suggest that "episodic reconstruction is just an adaptive design feature of the future planning system" (p. 393).

Recent memory studies have provided evidence to support the view that the constructive nature of memory has adaptive value in that it contributes to future thinking. First, it is reported that amnesia patients also have difficulties in imagining the future (Hassabis, Kumaran, Vann, & Maguire, 2007; Klein, Loftus, &

Kihlstrom, 2002; Tulving, 1985b, pp. 4-5). Second, functional magnetic resonance imaging (fMRI) studies have revealed that there are significant overlaps between the brain activities of remembering the past and of imagining the future (Addis, Wong, & Schacter, 2007; Szpunar, Watson, & McDermott, 2007). The default mode network—including medial-temporal lobe, precuneus, posterior cingulate, retrosplenial cortex, and the temporo-parietal junction (Raichle et al., 2001)—is considered to underlie both remembering and imagining (Buckner & Carroll, 2007; Spreng, Mar, & Kim, 2009).

The evidence supporting the constructive episodic simulation hypothesis stands indirectly against the traditional idea of memory distortion:

> While it is tempting to conclude that memory distortions point to fundamental flaws in the nature or composition of memory, a growing number of researchers have argued that, to the contrary, many memory distortions reflect the operation of adaptive processes—that is, processes that contribute to the efficient functioning of memory, but as a consequence of serving that role, also produce distortions. (Schacter, Guerin, & St. Jacques, 2011, p. 467)

Such a perspective suggests a new view on the function of memory.

As we will see later in §6.2, memory does many things; however, what is the proper function of memory processing system?[8] That is, what is the thing memory *ought to* do? As we have seen in both the previous and the present sections, if the function of memory is regarded as that of re-presenting past experiences then its constructive nature seem to be defective. Further, such a view cannot account for the adaptability of constructive memory. The idea that memory contributes to future simulation is motivation for a reconsideration of the function of memory. The constructive episodic simulation hypothesis (Schacter & Addis, 2007a, 2007b), by linking episodic memory and future thinking, provides an account of how memory can be put to use for current and future goals. This suggests that the constructive processing of memory, instead of being the operation of a defective system, is an adaptive process (Schacter, 2012; Schacter et al., 2011) which flexibly combines the

[8] "Proper function" proposed by Ruth Garrett Millikan (1984) refers to what the mechanism, trait, or process in question is *supposed to* do. It is a normative concept in the sense that is it not a causal or dispositional notion.

elements from past experience in order to simulate future episodes. The advantage of future simulation outweighs the cost of memory distortion.[9]

Suddendorf and Corballis (2007) suggest that the function of memory and the anticipatory system is to provide behavioral flexibility and to examine the phylogenetic development of different memory systems. According to them, the anticipatory behavior that different memory systems can offer vary in degree (p. 301, Figure 1). The most primitive one is procedural memory. It enables stimulus-driven predictions of regularities and allows behavior to be modulated by experience. The resulting behavior is, consequently, stimulus-bound. Declarative memory provides more flexibility, because they not only can be retrieved involuntarily, but also can be voluntarily triggered top-down from the frontal lobe, which enables decoupled representations that are not directly tied to the perceptual system. Such decoupled representations are only available when the constraint of offline activation is satisfied (see §2.2.3). In declarative memory, semantic memory is more primitive than episodic memory as it has less scope for flexibility. Semantic memory, in allowing learning in one context to be voluntarily transferred to another, provides the basis for reasoning. However, this is about regularities and not particularities. Episodic memory supplements this weakness: Through mental reconstruction or memory construction, it not only recreates past events, but it also allows the learned elements to be incorporated and arranged in a particular way in order to simulate possible futures. It thereby provides greater flexibility in novel situations and provides for the possibility of making long-term plans, extending even beyond the life span of the individual.

Furthermore, the view that memory increases one's behavioral flexibility is also supported by a recent view that memory may also play the role of prediction (Bar, 2009; Bar & Neta, 2008; Davachi & DuBrow, 2015; Lin, 2015; Lisman & Redish, 2009). Moshe Bar (2009) suggests that "our perception of the environment relies on memory as much as it does on incoming information" (p. 1235). Since we seldom encounter completely novel objects or events, our systems heavily rely on representations stored in memory systems to generate predictions. According to Bar's "analogy-association-prediction" framework (Bar & Neta, 2008), once there is a sensory input, the brain actively generates top-down guesses in order to figure out what that input looks like (analogy); the match triggers the activation of the associated representations (association), which allows predictions of what is likely

[9] According to this new view of memory function, memory distortion is not necessarily negative; instead, it can be beneficial to the organism.

to happen in the relevant context and environment (prediction). That is, brains proactively compare incoming signals with existing information gained in the past (see Bar 2009, Figure 1 & Figure 2).

Felipe De Brigard (2013) further investigates the function of remembering. In his view, both future simulations and past counterfactual simulations are to be treated as the same system of memory. He begins by raising the question of why we have such a constantly and systematically malfunctioning cognitive system. He goes on to argue that misremembering should not be treated as memory malfunction but as the *normal* result of a larger cognitive system in which remembering is just one part of the operation. This larger cognitive system supports the activity of "episodic hypothetical thinking", which he defines as "self-centered mental simulations about possible events that we think may happen or may have happened to ourselves" (p. 19). That is, it allows us not only to think of "what *was* the case and what potentially *could be* the case, but also what *could have been* the case" (p. 4). This idea finds support in a recent study (De Brigard et al., 2013) that shows that episodic counterfactual thinking (especially realistic thinking) involves similar brain regions to episodic memory.

De Brigard endorses Carl F. Craver's (2001) idea of a "mechanistic role function", which describes an item's function "in terms of the properties or activities by virtue of which it contributes to the working of a containing mechanism, and in terms of the mechanistic organization by which it makes that contribution" (p. 61). Based on this view of function, to determine the function of a system one has to determine the mechanisms of lower level activity, and how they contribute to higher-level functioning. Therefore, to determine the mechanistic function of memory requires an investigation into the way that its components contribute to the system and then how memory contributes to the functioning of the organism, allowing it to reach its goals (De Brigard, 2013). By arguing that certain misremembering also satisfies the requirements for memory construction as remembering does, De Brigard claims that to distinguish function from malfunction is unrelated to remembering and misremembering. Furthermore, considering the contribution to the higher level, he suggests that memory belongs to a larger cognitive system of "episodic hypothetical thinking". Such a view dismisses the problem of explaining how a system constantly and systematically malfunctions that we find in the traditional view.

Therefore, the distinction between memory function and malfunction is not equivalent to the distinction between remembering and misremembering or veridical

representation and misrepresentation; instead, it is goal-directed: It depends on whether the functioning of memory systems successfully contributes to the goals of the organism. That is, certain memory misrepresentation can equally lead to successful meta-cognitive function, behaviors, or actions; furthermore, there are cases in which misrepresentation rather than veridical representation leads to achievement of the goal at the higher levels. This is supported by the study of false recognition in amnesia patients, in which Schacter and colleagues (1996) show that amnesic patients are less susceptible to false recognition in memory tests with the Deese-Roediger-McDermott paradigm. Thus, it is suggested that "failing to misremember" can be seen as an indication of a pathological memory system.

I agree with the view from Suddendorf and Corballis that the main function of memory is to provide flexibility, and the view from De Brigard that the function and malfunction of memory is relative to the goals of the higher levels. Accordingly, in which ways can memory contribute to goals of the higher levels? First, constructive memory allows the system to think about possible futures, and counterfactual situations. These not only enable the subject to plan for the future, but also to reason for the best outcome by simulating different counterfactual scenarios. It has been shown that future and counterfactual simulation not only allows for more effective ways of coping, but it also leads to mood-enhancing effects that result from simulating positive outcomes (Brown, Macleod, Tata, & Goddard, 2002; as cited in Schacter, Addis, & Buckner, 2008, p. 50). Furthermore, future simulation allows for the reduction of temporal discounting (Boyer, 2008; Peters & Büchel, 2010). This enables the subject to resist immediate reward and opt for a more distant one. Furthermore, future simulation is a contributing factor in social interaction, not only because episodic simulation allows one to predict the future through immediate benefits of cooperation, but it also contributes by providing social contexts and facilitating social decision-making (Klein et al., 2009).

What is most important and what will be the focus of the following chapter is the way in which memory contributes to our mental autobiography. We have an idea of what kind of persons we are, and we know our story through reflecting on ourselves, on what has happened to me in the past, what is about to happen, and what kind of general characteristics I have. These are sustained by memory. I will develop the concept of the ASM in the next chapter. It is a collection of self-simulations of the relation with past and potential future states. These simulated episodes are structurally connected in a way that is isomorphic to our personal

history. It not only allows us to have self-understanding, but it also provides us with a sense of personal identity—the feeling that I am the same person as the one that existed in the past. I will discuss these issues in detail in §3.

2.4 Summary

This section aims to lay a theoretical background to memory and for discussions in later chapters. It includes the delineation of concepts of memory, the nature of memory, the mechanism and functions of memory.

First, concerning the concept of memory, I have reviewed different use of the term, taxonomy of memory systems, the differences between episodic and semantic memory, and the concept of working and autobiographical memory:

- *The term "memory" can be utilized to refer to different aspects of the memory process, including neurocognitive capacity, hypothetical store, information persisted, retrieval of information, and phenomenal awareness of remembering. In this dissertation, without any specification, I refer to the retrieved information (see §2.1.1).*
- *I adopt the taxonomy (see Figure 1) which classifies memory into different systems: long-term memory is differentiated from short-term memory; long-term memory includes procedural and declarative memory; the latter can be further differentiated into semantic and episodic memory (see Table 1 and §2.1.4 for the distinction between them).*
- *Working memory provides an interface for long-term memory to be integrated with sensory inputs and to be manipulated (see §2.1.3 for Baddeley's model of working memory). It makes a phenomenal self-model possible by providing a workspace, holding phenomenal representations temporarily active (see more in §3.2.1).*
- *Autobiographical memory is different from episodic and semantic in that it refers to the retrieval of information which is self-related, rather than memory systems. It is the outcome of the interplay between the episodic and the semantic memory systems (see §3.2 for ASM).*

Second, a representationalist account of memory is developed within a modern version of representationalism, which rejects the understanding of representations as mental entities. As we will also see in §3, the discussions in this

dissertation are based on a representationalist and functional account: the representationalism of memory and the self-model theory (see §3.1). The critical concept of an ASM is built upon them.

- *According to this account, the concept of representation can be analyzed in terms of a tripartite relationship between representandum (the object of representation), representatum (the internal state that carries information about the object), and representation (the process by which the system generates the representatum to create an internal depiction of the representandum) (see §2.2.2). In contrast to a representation which is the process that generates an internal state of the current state, a simulation is an information process that generates a simulatum, which simulates a counterfactual state of the external world for the (bio)system. Memory is considered simulata which simulates the past state of the external world (or the internal state of the system).*
- *Self-simulata, differentiated from simulata which, are possible selves: They are respectively utilized by (bio-)systems in order to simulate counterfactual states about themselves and the world to achieve the goals. Such a distinction characterizes the distinction of self- and world-related memory (see §2.3.2).*

Last, I look into the constructive nature of memory and its function. Based on the constructive nature of memory, three functional constraints for being an ASM are developed in the next chapter (in §3.2.5). Besides, this nature of memory has led to the recent reconsideration of the function of memory, which will play an important role in determining the distinction between memory enhancement and treatment (see §4 and §6).

- *Recollection is not the retrieval of intactly stored memory; instead, it is a constructed upon retrieval with a process of pattern completion, which allows the features to form a past experience (see §2.3.1).*
- *Newly developed views have regarded the function of memory as increasing the behavioral flexibility of an organism and adaptability to the natural and social environment. It allows the organism to simulate counterfactual scenarios and to act differently in accordance with these*

(see §2.3.1). The other important function of memory is that it allows us to form an ASM (see more in §3.2.1). We can, consequently, have a mental autobiography, informing us of what kind of persons we are.

Chapter 3

Conceptual Tools II:
Selfhood, Personhood, and Identity

3.0 Introduction
3.1 Self and Self-Consciousness
- *3.1.1 The Problems of the Self-Consciousness and Phenomenal Self*
- *3.1.2 The Nature of the Self*
- *3.1.3 Thomas Metzinger on the Self-Model Theory of Subjectivity*
- *3.1.4 The Phenomenal Self-Model*
- *3.1.5 The Phenomenal Subjectivity*

3.2 The Autobiographical Self-Model and Memory
- *3.2.1 Memory and the Self-Model*
- *3.2.2 Antonio Damasio on the Autobiographical Self*
- *3.2.3 Todd E. Feinberg on the Ego Dysequilibrium Theory*
- *3.2.4 The Autobiographical Self-Model*
- *3.2.5 The Functional Constraints for an Autobiographical Self-Model*

3.3 What I Am
- *3.3.1 Being a Human Animal: The Biological Approach*
- *3.3.2 Being a Person: The Psychological Approach*
- *3.3.3 The Moral Significance of Person and Selfhood*
- *3.3.4 A Normative Concept of Person*

3.4 Transtemporal Identity
- *3.4.1 The Concept of Identity*
- *3.4.2 The Sense of Identity*
- *3.4.3 The Biological Approach and Transtemporal Human Identity*
- *3.4.4 The Psychological Approach and Transtemporal Personal Identity*
- *3.4.5 The Moral Significance of Transtemporal Identity*

3.5 Summary

3.0 Introduction

The issue of authenticity, as I will later discuss in detail in §7-§8, is one of the main ethical issues in the cognitive enhancement (CE) debate. The dispute between critics and proponents of CE not only results from their different interpretations of the concept of authenticity but also from the different accounts of self, person, and identity they endorse. The former worry that enhancement technology threatens our "true self" and leads to loss of identity. This concern is based on a background assumption that there is a static self that is essential to us, and it remains constant across time. The latter, on the other hand, endorse another idea of self, and hold a variant idea of what one essentially is. In order to delineate this debate, the current chapter aims to provide the conceptual tools for a clear understanding of the concepts of self, person, and identity. To conclude, I will provide a general picture of the relation between memory and the concepts of self, person, and identity, based on the *Self-Model Theory of Subjectivity* (SMT; Metzinger, 2004).

I first focus on the problems of the self (§3.1). The SMT provides a comprehensive representational and functional theory to account for self-related phenomena: Self-models, together with functional constraints, serve to explicate different states of self-consciousness, and give a picture of the nature of the phenomenal self and subjectivity. Then, turning to the narrative aspect, which is linked to memory, after reviewing the idea of the *autobiographical self* (Damasio, 1999, 2010) and the *Ego Dysequilibrium Theory* (Feinberg, 2009a, 2009b), based on the SMT, I introduce the concept of an *autobiographical self-model* (ASM) in §3.2.4 and three functional constraints (§3.2.5). The third section (§3.3) deals with the question of what I fundamentally am. The two rival claims, from animalism and the psychological approach, are discussed, and a normative concept of the person is introduced. Last, the issue of transtemporal identity is addressed. Both the ideas of human identity and of personal identity are discussed (§3.4).

3.1 Self and Self-Consciousness

While there are variant understandings of the "self", and the problem of the self has a long history, the concepts remain elusive. This section aims to point out the target phenomena and introduce a functional and representational framework—the SMT proposed by Thomas Metzinger (2004)—to account for them.

3.1.1 The Problems of the Self-Consciousness and Phenomenal Self

The complexity of the discussion surrounding the "self" partially results from the diversity of the understandings of the concept. In general, two concepts of the self are differentiated (e.g., Glannon, 2007) distinguishes: One—the narrower concept of the self—targets the phenomenal experience of being an experiencer (Metzinger, 2004; Strawson, 1999a); the other—a richer concept of the self—includes a variety of self-related phenomena (Ramachandran, 2003; C. Taylor, 1989). The second concept covers the defining characteristics suggested by Vilayanur S. Ramachandran: continuity, unity or coherence, the sense of embodiment or ownership, and the sense of agency. Glannon (2007) adds a further component—"the ability to perceive and respond appropriately to the external world" (p. 33). Charles Taylor (1989), on the other hand, in order to develop a moral framework, holds a concept of the self that refers to the narrative dimension—our assessment about the world and ourselves (pp. 51–52).

What differentiates these two concepts is that the latter considers the qualitative aspect of the self, whereas the former aims to tackle the core mechanisms that allow for the emergence of these diverse phenomena. In this section, I will focus on the target phenomena of the first concept—the phenomenal self (Metzinger, 2004, p. 5). Then, in the following sections, I endorse and introduce the SMT. Based on the SMT, in §3.2, I develop the concept of the ASM, which is subsumed in the richer concept distinguished above.

Despite different views on the concept and the nature of the self, there remains a consensus that there are different kinds of phenomenology associated. We, as the experiencers of them, cannot deny their existence: Once we wake up every day, we feel, we experience, we think. Every time these events occur, they are experienced as if there is a feeler, an experiencer, or a thinker. It seems that there always exists a subject who undergoes such events and owns the feeling, the experience, and the thought. What is the nature of this subject? One may answer: It is I! It is "I" who is undergo all of these. But what am I? As René Descartes (1985, p. 18) asked, "I know that I exist; the question is, what is this 'I' that I know?" Likewise, Thomas Metzinger (2009) also poses the question in the beginning of his book, *The Ego Tunnel*:

> Why is there always someone *having* the experience? Who are the feeler of your feelings and the dreamer of your dreams? Who is the agent doing the

> doing, and what is the entity thinking your thoughts? Why is your conscious reality *your* conscious reality? (pp. 1-2)

This is how most of us experience: We experience in such way that there is a being, an "I", which is the "center" of *my* world and the recipient of all my sensations.

This *experienced* self is called "phenomenal self" by Metzinger (2004, p. 5). He differentiates three kinds of phenomenal experience: "mineness", "selfhood", and "perspectivalness" (2004, pp. 302-304; 2008, p. 217). *Mineness* is a higher-order property of particular forms of phenomenal content. It is the conscious experience that we own something, such as "my body", "my thought", and "my volitional act", used in our everyday expressions. *Selfhood* refers to the subjective feeling of "being someone". This is closely related to perspectivalness, which is "the structural feature of phenomenal space as a whole". This property is the first-person perspective in the sense that, phenomenally, I am the subject at the center of my world.

In addition to these two experiences, another fascinating aspect of our self-consciousness is that we can reflect upon ourselves. We not only think as a thinker, but can also think about ourselves as the one who thinks. In everyday life, we interact with the world and with ourselves; these interactions include perceptions, intentions, and actions. The capacity to reflect means that such interactions can be represented in our conscious experiences: There is a world, a self, and a relation between them; that is, we not only experience but also experience as *being a self experiencing the world.* This is *perspectivalness*.

What's more, when we reflect upon ourselves or question the nature of ourselves, we not only come up with the idea that we are subjects of our experiences, but also answer with some properties that distinguish ourselves from others (e.g., female, outgoing, enjoy music). We know where we are from, what kind of person we are, what we like and dislike, and what we have experienced and expect to go through. It is our personal history, personality, habits, and other characteristics that differentiate us from others.[10] First, we no longer live only in the present: We have a history of our own. We can recall what has happened in the past. We also have a future, with purposes: As we "move toward the future", our current desires and actions are for the sake of future goals. This gives us a strong intuition that we have lived in the past and will live in the future, and accordingly results in our strong

[10] This is the difference between the question of "who I am" and "what I am". The former considers the content, while the latter questions what enables the realization of such contents.

interests in the question of personal identity: Am I the same person as the one that existed yesterday? We will go deeper into this issue in §3.4. Second, the autobiographical aspect not only makes us unique in the social world, but also provides us with meaning: We have preference and we can value. We not only know our relation to environments or other social beings, we attribute values to them. Therefore, we have friendships, relationships, a hometown, and other meaningful relations. This is what we will focus on in §3.2.

I have only illustrated the most relevant phenomena for our discussion of memory enhancement. There are other "self-related phenomena". For instance, what philosophers refer to as "bodily self-consciousness" aims to investigate how bodily sensation contributes to our phenomenal experience. This cannot be ignored by a comprehensive theory of self-consciousness.

3.1.2 The Nature of the Self

As we will discover in the last part of the dissertation (§7–§8), to be authentic is understood as being true to one's self, and enhancement technologies are worrisome for critics of CE, because they may threaten one's existence by making one lose one's self. Such a view presupposes that there exists a static self, without which we no longer exist. This section considers the metaphysical nature of the self and different accounts of the metaphysics of the self.

Take the metaphysical questions of the self: Does there exist such a thing as the self? Does the self exist as a substance or an entity? If the answer is affirmative, what kind of entity are we considering? Is it a physical entity, a psychological entity, or an entity with both physical and psychological properties (Bermúdez, 2008, pp. 457-458)? If there is no such a thing as the self, how do we explain self-consciousness and phenomenal self? And how do we understand our strong intuition that there is such a self? These questions constitute the debate between realism and anti-realism of the self.

Ontological realists of the self contend that the self exists as a substance, the existence of which is independent of anything else. For instance, Martine Nida-Rümelin (2010) argues for a version of substance dualism through the problem of personal identity. She contends that "there is an individual that has experiences, thinks and is active and is *neither identical* to any material thing *nor constituted* by any material thing" (p. 191). John Foster (1991) objects to different versions of reductionism, and claims that there exists a non-physical subject that owns mental states and is *basic* in the sense that it is conceptually and metaphysically fundamental (p.

203). However, such theories of the self presuppose an ontology of the world that is beyond current scientific explanations. No empirical study has directly or indirectly shown the ontological existence of a self, and no conceptual argument has suggested the necessity of such assumption.

Ontological anti-realism about the self denies that the self exists as a substance, which is one of the fundamentally ontological entities. David Hume (2004), for instance, denies the ontological existence of the self, and proposes a bundle theory, according to which we are merely a bundle or collection of perceptions (p. 180). Daniel Dennett (1992) further defines a self as a "center of narrative gravity", or merely a fictional object. Buddhist reductionism (Siderits, 2011) contends that the self—the essential part of the psychophysical complex—does not exist, whereas the person—the psychophysical complex as a whole—is a conceptual fiction (p. 298).

3.1.3 Thomas Metzinger on the Self-Model Theory of Subjectivity

The SMT proposed by Metzinger (2004) is a naturalized theory of consciousness, selfhood, and subjectivity. The SMT analyzes phenomenal properties with representational and functional properties, and shows how an information-processing system is considered a phenomenal self by satisfying a series of functional constraints. Thus, according to SMT, "there is no such things as selves exist[ing] in the world: Nobody ever *was* or *had* a self" (Metzinger, 2004, p. 1). That is, what folk psychology means by "the self" is not a substance or an unchangeable essence, which exists independently, and if it is, then it does not exist. Instead, what exist are the experience of being a self and the contents of self-consciousness, which change constantly (Metzinger, 2008, p. 215). And, as we will review later, this *experienced self* or *ego* is the content of our representations.

To begin with, SMT searches for the minimally sufficient condition for a conscious self. Such an approach, of looking for an elementary form of self-consciousness, is also adopted by Galen Strawson (1999b) to delimit the minimal self. Nevertheless, unlike Strawson, who holds that phenomenology can constrain metaphysics, Metzinger holds that our phenomenology, e.g., the phenomenology of substantiality (Metzinger, 2011, p. 283) does not suggest any metaphysical realism of the self. His SMT aims to provide a representationalist and functionalist account of the minimal phenomenal self: What is the set of representational and functional properties that a system requires in order for a minimally phenomenal self to emerge?

First of all, the system requires a *self-model*—a coherent self-representation, a consistent internal model of itself as a whole (Metzinger, 2008, p. 218). A self-model is "a model of the very representational system that is currently activating it within itself" (2004, p. 302). It is an information-processing system that not only internally and continuously *simulates* its own observable output but also *emulates* abstract properties of its own internal information-processing (Metzinger, 2004, pp. 300-301). The concept of emulation is distinguished from that of simulation: To simulate is to represent the properties accessible to sensory processing and their development over time; to emulate is to internally simulate not only the behavior of the target system but also its hidden aspects. A self-model is the special model whose "target system and simulating/emulating system are identical" (p. 301).

Such a self-representing system is not conscious. The motor program integrated into the unconscious self-model is self-directed behavior, and only after it is integrated into a *phenomenal self-model* (PSM) does self-directed behavior turn into self-directed action. That is, it can be consciously aware only if the information is integrated into a PSM.

3.1.4 The Phenomenal Self-Model

How does an unconscious self-model become a PSM? According to SMT (2004), *globality*, *presentationality*, and *transparency* are the three minimal constraint satisfaction for consciousness (p. 204). First, *global availability*. To satisfy this constraint, a functionally active model emerges within the system: Information becomes available to the system, and there is a process of generating a coherent and constantly updating model of itself as a whole. This results in the availability of the contents of conscious experience to cognitive capacities, such as attention, cognition, autobiographical memory, speech, and action control. If the constraint of globality is satisfied, the contents of my phenomenal self-consciousness are directly available to me in the sense that, phenomenally, I have a (graded) subjective feeling of immediacy that they are given to me (Metzinger, 2005, pp. 305-307). In addition, a coherent self-representation allows the system to have a self-world border, and enables an internal image of itself as a whole, distinct from others. Thus, system-related information (integrated into the self-model) is available to the system as self-related, and information that is not system-related (not integrated into the self-model) is considered non-self. In this sense, subjectivity and objectivity emerge simultaneously (pp. 307-308).

Second, we have *presentationality*. Presentationality is unique to phenomenal content: You cannot find the time "now" in a physical description of the universe; that is, "the physical world is 'nowless,' as well as futureless and pastless" (Metzinger, 2004, p. 127). "Now" only exists in phenomenal content (p. 126). No matter what you experience, you always experience at present. This is simulational content which requires no epistemological justification from reality: It is presented to the system as reality (p. 128). According to SMT, the main function of presentationality is to integrate the basic representational content that is currently carried out by bio-regulatory information-processing and to stabilize bodily condition, even while higher-order cognitive contents simulate possible states of the system. (p. 312).

Finally yet importantly, transparency is the most important constraint for a system to be considered a conscious system model[11]. According to the definition from SMT, "[t]ransparency holds if earlier processing stages are unavailable for attentional processing" (2005, p. 11). This functional constraint accounts for "the phenomenology of naïve realism" (2004, p. 170): It is as if I "see through" my internal representation and phenomenally I interact and contact with the environment directly (2007). Transparency is not an all-or-none phenomenon: Sometimes it can be opaque, when the world appears "unreal" or "dreamlike" (2004, p. 172). Information is presented to a transparent model as being factual, but when the system reaches a certain degree of opacity, the "appearance-reality distinction" becomes available to the system. When the constraint of transparency is applied to consider the self-model, the process of self-representation, which is neural interaction in the brain, is no longer available, and then one can merely and directly "see" the content of self-representation (p. 334). In this way, we become naive realists about ourselves (p. 339), and the conscious experience of "selfhood" is brought about.

So far, a PSM—a self-representing system satisfying the constraints of global availability, presentationality, and transparency—only allows the phenomenal experience of "selfhood" and "mineness" to appear. Such a system, merely equipped with minimal self-consciousness, is "frozen in an eternal Now" and the world appearing to it lacks any internal structure (p. 559). Besides, under such conditions, there is no "first-person perspective": This phenomenology has no structural feature of the subject being the center of the experience.

[11] "Transparency" refers to "phenomenal transparency" rather than epistemic transparency (Metzinger, 2003).

3.1.5 The Phenomenal Subjectivity

How does phenomenal subjectivity emerge? The other functional constraints—*convolved holism*, *dynamicity*, and *perspectivalness*—enable further complexity of the mental content and, most importantly, a phenomenal first-person perspective. As I will explain, this involves another conceptual tool—the *phenomenal model of the intentionality relation* (PMIR).

First, the functional constraint of convolved holism accounts for the existence of internal structures. One example of phenomenal holism is "perceptual object-formation", which refers to the integration of features of a representational unit. The property of holism is convolved, dynamic, and flexible: The phenomenal self is a multitude of internal part-whole relationships and constantly undergoes changes (Metzinger, 2004, pp. 321-323). Convolved holism results from the coherence of system-related information represented to the system as a single unity, and from an "internal kind of interdependence" of integrated information, which is a way of representing the causal structure within the model (p. 324).

In addition to "presence" (accounted by the functional constraint of presentationality), the functional constraint of dynamicity characterizes the possibility of duration and change in phenomenal content; that is, we are living not only in the present, but also in a present with past and future (Metzinger, 2004, pp. 151-154). At a representational level, first, the individuation of events is represented, and next, the pattern of the sequence in time is formed; thus, we are able to experience the change of an object through time (object identity) (pp. 154-155). As "the temporal structure of our behavioral space" is represented, autobiographical memory and future thinking (such as planning) become possible (p. 155). The "flexibility of our behavioral repertoire" can be increased, as we are no longer just fixed to nowness. However, it is worth noting that the constraint of presentationality still applies: Even when we recall memories, plan for the future, or engage in other simulational content, we are doing these "now" and never lose touch with the present; the nowness is felt as "being present as a self", and is situated in a temporal order that includes the past and the future.

Finally, in order to possess a first-person perspective, the PMIR is introduced. PMIR is an inner mental representation of the subject-object relation: When we are directed towards an object (e.g., perceiving an object), not only the object itself but also the representational relation is represented. This creates "a self in the act of knowing" or "a self as intending to act" in the content. That is, the system's "phenomenal space is a *perspectival* space, and its experiences are

subjective experiences" (Metzinger, 2008, p. 241). The object can be a perceptual object, action goal, or another self. The representation of the subject-object relation makes new kinds of information available to the system and new mental capacities, such as agency (being aware of itself as an agent with a will), attentional subjectivity (having high-level selective attention), reflexive self-consciousness (being directed at the PMIR itself), other agent-modeling (being aware that other systems have a first-person perspective), mind-reading (internally simulating other systems' PMIRs), and high-level intersubjectivity (acknowledging other systems as persons) (2005, pp. 29-30).

Another important constraint related to memory (episodic or autobiographical memory, to be more specific) is *offline activation*. As I have showed in §2.2, the distinction between presentation, representation, and simulation from Metzinger (2004) illustrates the difference between simulation and the other two, namely presentation and representation: Simulation is generated largely independently of current sensory input, and the content of phenomenal simulation does not refer to the current actual state of the world. The constraint of offline activation considers the emergence such "representation". Dreaming is an example of a global offline state (pp. 251-264): Dream states are world-models resulting from the brain's continuous interpretation of internal stimuli. Due to the capacity to generate globally available offline simulations, a new kind of information is available to the system: The system can create a possible world or a possible self (e.g., past, future, or imagined states of the world or itself), and can differentiate the real from the possible world or self. This happens when we recall episodic memory, plan for the future, daydream, or let our minds wander. Simulation and self-simulation allow the organism to internally "rehearse" not only possible future scenarios but also potential perceptual perspectives, sensory states, and actions (pp. 182-183).

3.2 The Autobiographical Self-Model and Memory

This section aims to delineate the relationship between memory, the self-model, and the ASM. We have a mental autobiography or narrative that allows us to know our personal history and what kind of person we are. This is made possible because we can integrate our ASM into the PSM. I will first discuss how self-representation and functional constraints are related to memory. Then I will review the idea of the autobiographical self from Antonio Damasio (1999, 2010) and the ego

disequilibrium theory of Todd E. Feinberg (2009b). Later, based on the previous discussion and the SMT (Metzinger, 2004), I will illustrate the concept of an ASM and introduce three functional constraints for a self-model to be considered an ASM.

3.2.1 Memory and the Self-Model

In the last chapter (§2.1) we looked at different kinds of memory: Long-term memory can be differentiated into declarative and non-declarative memory. Among the former, episodic memory and semantic memory are further distinguished. Besides, working memory is an important short-term memory system that is closely related to the formation of PSM. This section illustrates the distinct relations between the self-model and these kinds of memory.

Let's start with working memory. Working memory is necessary for the emergence of a PSM. With four components—central executive, episodic buffer, visuo-spatial sketch-pad, and phonological loop—it is a limited capacity for storage that allows the system to manipulate information in order for it to be used for further cognitive processing (Baddeley, 1983; 2000a; §2.1). Working memory makes a PSM possible by providing a workspace, which holds the phenomenal representations temporarily active. The central executive binds information from different resources into a coherent episode, whereas the episodic buffer stores information temporarily and serves as an interface for different information resources, i.e., visuospatial sketchpad, phonological loop, and long-term memory. The integrated information is the content of PSM and is accessible to consciousness.

Further, episodic and semantic memory are the cognitive systems that enable the mental (self-)simulation, which is structurally isomorphic to the phenomenal representation that was activated in the past, and which is integrated into the conscious self-model and becomes consciously and cognitively accessible. PSM is a necessary condition for the successful functioning of declarative memory. In order to turn mental representations of one's past self—or world-model—into phenomenal representation, a PSM is required for the mental representations to become consciously and cognitively accessible.

Furthermore, more functional constraints are involved: *offline activation*, *dynamicity*, and *perspectivalness*. First, as discussed in §2.2, unlike perception, which refers to the actual state of the world, *offline activation* allows the system to construct a counterfactual mental model that is disconnected from the current situation and goes towards forming a possible self. This, therefore, allows a simulation of the self-model that was once active. Second, the constraint of

dynamicity characterizes a system that is not only tied to "nowness" but also own a personal history. Consequently, we constantly experience as if we leave the past and move forward to the future.[12] Episodic memory is not possible without this aspect, for one important feature of episodic memory is that the content always refers to the past, whether the exact point in time is clear to the individual or not. Last, the constraint of perspectivalness (PMIR; see §3.1.5) considers the representation of subject-object relations and the ability to recognize oneself as a subject. As such, the current model cannot only integrate the simulated self-model but also form a relation with it. As I will explain in more detail in §3.4.2, such a relation allows the phenomenal experience whereby I identify the past subject as the same being, and that I have continuously existed from the past until the present moment.

It is worth noting that a self-world boundary that is consciously available to the subject is required for both semantic and episodic memory: One critical but not decisive differentiating feature is that the former is world-related, while the latter is self-related (see §2.1 and Tulving, 2005). (Though semantic memory also contains personal information, it is represented in a way that it is part of world knowledge.) Without a self-world distinction, the present senses of episodic and semantic memory are not possible. This thus implicates the discussion of the memory of the animal I mentioned in §2.1: Which kind of declarative memory is more primitive, episodic memory or semantic memory? Do animals have episodic memory? Many (e.g., Suddendorf & Corballis, 1997, 2007; Tulving, 2005) have argued that episodic memory evolved more recently and is uniquely human, while semantic memory is more primitive and can be found in non-human animals. However, I am inclined to an alternative view: Those non-human animals without first-person perspective (PMIR) possess another kind of memory, distinct from both episodic and semantic memory. This primitive form of declarative memory is that from which episodic memory and semantic memory are developed once the system is capable of forming a PMIR and a self-world boundary is available to it. This kind of memory is considered neither self-related nor world-related, because there is no self-world distinction (or boundary) available to it.

As for most human beings equipped with both kinds of declarative memory, episodic and semantic memory cannot be neatly distinguished. Evidence has shown that semantic memory and episodic memory are interdependent and work closely together to enable a successful creation of phenomenal self-simulations that refer to

[12] The "past" and the "future" refer to the temporal concepts mentally constructed. It is distinct and can be dissociated from the physical concepts.

past PSMs (Greenberg & Verfaellie, 2010). Together, they enable the emergence of a mental autobiography or narrative. The next sections will focus on this aspect and introduce the concept of the ASM (§3.2.4).

3.2.2 Antonio Damasio on the Autobiographical Self

Damasio (1999, 2010) proposes the idea of an *autobiographical self*. It will be helpful to give a brief review of his theory of self and the distinctions of *proto-self*, *core self*, and *autobiographical self*. According to Damasio (1999), the self is central to the understanding of human conscious mind. What is "self"? According to Damasio, there is self, but it is not a thing: Self is the process which presents whenever we are conscious (Damasio, 2010, p. 8). Unlike the SMT (Metzinger, 2004), which claims that there exists no such thing as *the self* except for the phenomenal self, which is the content of our experience, Damasio contends that the most primitive form of self-process starts to appear after the establishment of the mind and alertness, and before the emergence of consciousness. Consciousness is generated only after the more complex level of self is developed. In light of different stages of the self-processes, he distinguishes three kinds of self: proto-self, core self, and autobiographical self (Damasio, 2010, p. 10).

Prior to the emergence of consciousness, as a preconscious biological precedent the proto-self is a coherent collection of neural patterns which involves in the representation of the current bodily state (Damasio, 1999, p. 154). The proto-self is the "felt body image" resulting from the coordination of three kinds of maps (2010, pp. 190-199): the "master interoceptive maps", composed of signals from the internal milieu and viscera, report the internal state of the organism to the central nervous system; the "master organism maps", on the other hand, describe a schema of the body in repose, including the major components of the body such as head, trunk and limb; the "maps of the externally directed sensory portal", with frontal eye fields and somatosensory cortices both involved, build perspective and feelings such as "sense organ location". Working with each other, these maps, moment by moment, contribute to the proto-self, as the most stable aspect of the organism's physical structure and the stepping-stone required for the emergence of the core and autobiographical selves. As the first-order representation of current body states, it acts as a platform for consciousness or the "something-to-which-knowing-is-attributed" (1999, p. 159).

The core self exists in a second-order nonverbal account, which captures the causal relationship between object and organism. Equipped with the core self, the

organism can internally construct and exhibit a kind of wordless, salient knowledge of the changes in the organism induced by an object. By re-representing the relationship of the proto-self and the object, the organism is able to be represented in such way: "proto-self at the inaugural instant; object coming into sensory representation; changing of inaugural proto-self into proto-self modified by object" (Damasio, 2010, p. 177). It is when the organism begins to generate subjectivity that it thereby qualifies for consciousness (p. 182). The core self, making the organism-object relationship available, allows the organism to ask and answer the question that has never been posed before: What is happening, and what is the relationship between the organism and objects? The acquisition of the core self is the beginning of the opportunity not only to comprehend a situation but also to plan responses. It is at the same time accompanied by an enhancement in the degree of wakefulness and attention (1999, pp. 182-183). The development of the core self is close to the PMIR of the SMT (see §3.1.5). Not only the organism itself but also its relation with objects is represented. It allows the *subjective* experience to emerge.

The autobiographical self hinges on the same "you" as the core self, but now the "you" is connected with the "you" of another time, and together these construct the parts of the autobiographical record (Damasio, 1999, p. 195). "Autobiographical memories are objects, and the brain treats them as such, allows each of them to related to the organism in the manner described for core consciousness, and thus allows each of them to generate a pulse of core consciousness, a sense of self knowing" (pp. 196-197). Memories reactivated act as "something-to-be-known" triggering the mechanism of core self, and result in the autobiographical self. In order to go beyond the core self and to acquire the autobiographical self, one must first have the ability to retain records of experience from the older core self and then the ability to reactivate these records to generate "a sense of self knowing", which must be held active for a substantial amount of time. As such, there is a close connection between memory and autobiographical self. Memories contained in brain networks are dispositional and implicit, and can be simultaneously made explicit at any time (p. 174). Autobiographies are made up of personal memories, namely the sum of what we have experienced, planned, and expected in our lives. The autobiographical self is autobiography made conscious. The autobiographical self is, therefore, based on autobiographical memory, and is constituted of implicit memories of an individual's past and future. The autobiographical self extends overtime as autobiographical memory grows through a lifetime, with the accumulation of the experiences. The autobiographical self is also apt to be partially

modified and remodeled, as autobiographical memories are created from more recent experiences (p. 174).

3.2.3 Todd E. Feinberg on the Ego Dysequilibrium Theory

Todd E. Feinberg (2001, 2009a, 2009b, 2011) focuses on the self of human adults with a view to a better understanding of how the content of the self is organized. As a psychiatrist, Feinberg builds his theory of self on studies of self-related neuropathology. Feinberg defines the self by its coherence: "The self is a unity of consciousness in perception and action that persists in time. […] the self is something that at least subjectively feels like it endures beyond the passing moment" (Feinberg, 2009b, p. XI). It is not only what it is like to be me that I experience every day but also the collection of "me"s in a temporal sense. The idea of self Feinberg has in mind is similar to Damasio's autobiographical self.

Feinberg (2011, p. 75) differentiates the neuropathology of the self into perturbations of the "bodily self," "relational self," and "narrative self". Disorders of the bodily self refer to the alteration of how one views the nature or boundary of one's physical being. For instance, patients with delusional anosognosia show delusional denial of paralysis, and patients with somatoparaphrenia may misidentify their body parts. Disturbances of the relational self affect the manner in which the individual interacts with objects and persons by modifying the personal significance of the relation between the world and herself or himself. Delusional misidentification syndrome (DMS) includes Capgras syndrome, which is under-related delusional misidentification, and Frégoli syndrome, which in contrast is over-related misidentification of the relation with a person, thing, or place. Disturbances of the narrative self affect the way the individual describes her personal past and present circumstances. In personal confabulation, a patient may distort an actual event in her or his life or create a fictitious narrative about herself.

Feinberg (2001, 2009a, 2009b) proposes the concept of "personal relatedness" to account for the perturbations of the bodily self, relational self, and narrative self. This is a feeling that connects us with the items around us, as illustrated by Feinberg (2001):

> The persons, places, objects, and events that one's self experiences are imbued with feeling—the feeling of how one relates to things in a personal sense. Our identities are built around this sense of relatedness. Personal relatedness provides the structure within which the self is anchored in the world. The self is a continuum of relationships. An individual's own body, spouse, and family

members are “ego-close.” They bear a particular personal relationship to the self, identity with the self; we care about these items, these events, these people, in particular ways. They are significant. The objects of the world, which for us have no personal significance, could be considered “ego-distant.” The impersonal world, the stranger on the street, is less likely to be imbued with any sense of personal significance. (p. 30)

The whole network, based on personal relatedness acting as a tag marking the relation one has with one’s body, people, and events, gives one an idea of what kind of person one is, and how one is related to the world. Furthermore, according to the Ego Dysequilibrium theory from Feinberg (2001, 2009a, 2009b), such personal relatedness is supported by an *ego equilibrium*. This ego equilibrium is created by the balance between right and left hemispheres of the brain. Any disturbance of the ego equilibrium leads to the neuropathology mentioned above. As the ego dysequilibrium theory suggests, right frontal damage leads to the creation of a two-way disturbance between the self and the environment. These disturbances specifically link to personal relatedness, which could lead to disorders of both under- and over-relatedness to the environment. “Without the mediation of right frontal regions that subserve certain self- and ego-related functions, patterns of personally significant incoming information may be disconnected from a feeling of familiarity or personal relatedness” (2009a, p. 103).

But what pulls everything together here? As we see, the problem of mental unity poses a challenge. To this Feinberg proposes the “nested hierarchy of the self”. Nested hierarchy differs from non-nested hierarchy in two aspects: First, in nested hierarchy, the elements composing the lower levels of the hierarchy are physically combined within higher levels to create a more complex whole; in non-nested hierarchy, the physical entities that compose different levels are physically independent. This results in the second difference: In non-nested hierarchy, the higher levels commend the lower levels by sending signals, whereas in nested hierarchy, the higher levels impose control on lower levels by constraint (Feinberg, 2001, pp. 127-131; 2005).

The nested hierarchy of the self shows that the neural self is comprised of three hierarchically arranged and interrelated systems: the “interoself system” for homeostatic internal processes, the exterosensorimotor system for the responsiveness to the external environment, and the integrative self system for the assimilation of the previous two systems (Feinberg, 2009b, pp. 148-155). There are many levels in the neural self, including the tectal level, the level of “reptilian

brain", the limbic level, the paralimbic level, and the heteromodal association cortex. As a nested hierarchy, lower levels are nested within higher levels, and each of them contributes to the self. In the light of these levels of neural organization, various forms of consciousness are developed, from the simplest at the reptilian level, to the more complicated higher consciousness.

How do higher levels of the neural self constrain lower levels? Feinberg (2005) proposes that the sensory and motor systems, which are part of the exterosensorimotor system, are respectively constrained by meaning and purpose. Meanings produced at successive hierarchical levels are conjointly represented to produce the meaning of higher levels, and this higher level of meaning produces a "top-down" constraint on the constituting elements. In the motor system, the purpose of action sits at the highest level, and works by bringing lower-level elements into action, creating unity. "The neurobiological self can be understood as a nested hierarchy of meaning and purpose" (Feinberg, 2001, p. 7).

It is worth mentioning that Feinberg (2009b, pp. 155-157) adopts Tulving's model of three systems of memory: anoetic consciousness, noetic consciousness and autonoetic consciousness. Anoetic consciousness is supported by procedural memory, and it allows one to perform actions without being able to verbally explain or remember. According to Feinberg, animals with reptilian levels of neural organization are equipped with this simplest form of consciousness. Next, noetic consciousness, with semantic memory, enables one to be aware of and to cognitively operate on objects and events in an abstract way. This kind of consciousness is possessed by animals with neural organizations including paralimbic cortices. It is also similar to the "core self" of Damasio. Last, with the acquisition of episodic memory, autonoetic consciousness allows the emergence of self-awareness, identity, autobiographical identity, and a sense of being in time. It is believed that autonoetic consciousness depends on the frontal lobe structure, which is involved in the entire neural organization.

3.2.4 The Autobiographical Self-Model

An ASM is a collection of mental self-simulations of the relations with past and potential future states. The content of such self-simulation neither correlates to the current stimuli nor refers to the actual state of the model itself (or the world). Instead, it represents the past self or the potential future self. The simulational content is not only a self in the act of experiencing something that one has experienced before, or that one is likely to experience in the future, but is also a

temporal and emotional relation between that self and the current experiencing subject, that is, the current active self-model.

We are conscious of the content of the ASM when it is integrated into the active PSM: Mental simulations become phenomenal simulations. Past and future episodes become explicit, and we seem able to "re-experience" the past experience or to have "mental time travel" to the past or future. Nevertheless, as I have discussed in §2.1, the terms "re-experience" and "mental time travel" fail to correctly characterize this phenomenon. Take an episodic recall as an example. We are still aware of the current environment and are still bound to "now". Therefore we don't really travel to another mental time, rather, we instantiate it in our *current* PSM. The constraint of presentationality illustrates our ties to the current moment: No matter what I experience, I experience it now, even if the content refers to the distant past or future (or to the time before our birth or death). Besides, this is not exactly "re-experiencing", because our experience will not be the same as it was before. For instance, among many differences, the transparency of the phenomenal representational content is lost: the content of self-simulation is typically opaque. In addition, the content will be modified as it is integrated into the current PSM: How it is modified is largely dependent on the current self-model (see §3.4.2).

As soon as the ASM is integrated into the current active PSM, our phenomenal content is enriched in the following respects: First, phenomenally not only do we experience ourselves as a self that lives at the present moment within a temporal dimension, but this temporal dimension is "marked" with a series of personal events in the past, as well as plans and expectations for the future; that is, the ASM allows us to have a personal mental autobiography which starts from the past and extends into the future. However, as we will see in what follows, such a relation between us and our mental autobiography is different from our reading an autobiography.

Second, as one "moves forward in time", the structure of the ASM changes accordingly. Such structural changes include the represented temporal relation between the past or future episodes in question and the current thinking or recalling subject: The distant future becomes the near future; the future indeterminate content becomes determinate and the past. This can be thought of as "an inner revolving stage". In some musicals, "Les Misérables" for instance, in order to portray the fleeing of the main character Jean Valjean, the actor runs on a revolving stage[13]: In

[13] A revolving stage is a mechanically controllable platform with multiple scenes built-in, within a theatre. It is used to speed up the changing of scene, and to create a show without the gaps that result

this way, different scenes (including other fixed actors and actresses) rotate and pass by the sight of audience while Jean Valjean remains at the center of the stage and under the spotlight. When the scene is not under the spotlight, it is in the dark and not visible to the audience.

If we think of ourselves as the main character of the musical, and under the spotlight is the content of PSM, metaphorically speaking the future comes to the present (under the spotlight) and becomes the past, while offline simulation allows us to review the past by bringing a past episode back to the present PSM (i.e., turning the revolving stage backwards to bring the old scene back to the spotlight), or preview the future by temporarily internally rehearsing the future (i.e., fast forwarding the stage to the future event). The upshot is that what enters our consciousness must be integrated into our conscious self-model, which is fixed to a spatial and temporal point (i.e., now). This integration with the ASM can be regarded as the installation of a revolving stage, which allows us to bring our past or future content into the current PSM. Before turning to the next point, it is worth noting that it is an *internal* revolving stage, because unlike a musical there is no audience independent from the mental system: There exists no one observer that observes what is represented (Dennett, 1991). Instead, the representational system, by satisfying some functional constraints, is considered a system capable of observing the environment and itself.

Third, unlike an autobiography, the content of the mental ASM is not fixed. As I discussed in §2.2, the memory system is not like the storehouse suggested by John Locke (2008), but is constructive in nature. Accordingly, the content resulting from the integration of the autobiographical and PSMs is dynamic and "interactive", in the sense that it relies heavily on the current active self-model. For instance, any autobiographical content that is inconsistent with the current self-model is unlikely to be integrated into it or to become conscious. In addition, the way it is integrated is also dependent on the current purpose or goal of the self-model. Self-deception might be one example: For instance, to prevent acknowledgement of an inconvenient truth, the system may integrate the autobiographical content in such a way as to modify the story so that it avoids awareness of the complete story (von Hippel & Trivers, 2011, pp. 9-10). Here the current emotion and the represented emotional content will also influence whether and how the autobiographical information is integrated. In this aspect, the metaphor of a revolving stage may be

---

from the switching of scenes. This kind of stage design has been used since the 17th century and is now notably used in the musical Les Misérables.

misleading. To fix the context, the scene is reconstructed every time, when it is rotated to the center to adjust the current plot.[14]

Fourth, the constraint of global availability is satisfied, as the autobiographical simulation is made available to other cognitive processes. For instance, the memory about past interaction with one person can improve one's capacity of simulating that person's mind. In addition, there have been more studies on the benefits of self-generated thoughts with self-related contents (or mind-wandering): It allows us to connect with our past and future self-model (see §3.4.2), to make more-successful long-term plans (Smallwood, Ruby, & Singer, 2013), and to improve social interaction (Ruby, Smallwood, Sackur, & Singer, 2013). More benefits are illustrated by Damasio (1999) as follows:

> The ability to create helpful artifacts; the ability to consider the mind of the other; the ability to sense the minds of the collective; the ability to suffer with pain as opposed to just feel pain and react to it; the ability to sense the possibility of death in the self and in the other; the ability to value life; the ability to construct a sense of good and of evil distinct from pleasure and pain; the ability to take into account the interests of the other and of the collective; the ability to sense beauty as opposed to just feeling pleasure, the ability to sense a discord of feelings and later a discord of abstract ideas, which is the source of the sense of truth. (p. 230)

The ASM allows a new state of the self-model to emerge.

Fifth, the most important aspect of the ASM is that it constantly informs us who we are. Its function is not just to provide information, but also to serve as a framework for being in the world. Phenomenally, the world is comprehended in a way that is "egocentric": Things (e.g., agents, objects, places, and events) appear to us with certain meanings and degrees of significance. As Feinberg suggests, in his ego dysequilibrium theory (§3.2.3), personal relatedness weaves a web that indicates how we comprehend the world. Likewise, the ASM serves as a map that guides us in how to interact with our environment.

Last, the ASM is closely linked to the notion of personality. Personality refers to a collection of psychological properties of an individual from a third-

[14] Another reason why the metaphor is misleading is that when the ASM is incorporated into the current PSM, beside the simulation, PSM still also represents the current situation, which allows us to be aware of our actual environment when we are thinking about the past or future.

person perspective. One's personality is determined by two major factors: the traits that are more or less determined when one is born; and one's ASM, which results from interactions with the environment and other agents (Damasio, 1999, p. 222). As we will see, a diachronically coherent ASM allows for a more consistent personality across time, whereas dramatically diachronically incoherent ASMs result in a constant shifting in personality (e.g., dissociative identity disorder (DID)).

3.2.5 The Functional Constraints for an Autobiographical Self-Model

How is one's ASM generated? How is it that some elements and some related episodes are retrieved, while others are not? What are the functional constraints for a self-model to be considered an ASM? This section aims to introduce three functional constraints—synchronic coherence, diachronic coherence, and global veridicality—which can also be used to characterize ASMs. They are based on the account of Martin A. Conway et al., who investigate the organization of representations of memory elements and the constraints for formation of autobiographical memory. Here I will introduce Conway's account and then suggest three functional constraints for an ASM.

According to Conway et al., these elements "stored" in long-term memory are not just a random collection or pool of past events and learned facts. Rather, they are grouped, classified, or associated in a certain way, so that given a cue (e.g., a word, a name, an event, or a fact), certain representations are activated while others are not, and some are activated earlier and more easily than others. Besides, a particular memory has its own internal structure. For instance, whenever you recall trips with friends, there seems to be a pattern to how it is presented. Numerous psychological studies have mapped organization of memory by comparing the response time of the retrieval of memories when confronted with different cues. Subjects are supplied with a cue (e.g., a word) and asked to respond as soon as an associated memory comes to mind.

Conway and Pleydell-Pearce (2000) propose a theoretical autobiographical memory framework—the *Self-Memory System* (SMS). The SMS is a conjunction of two main components: the *working self* and an *autobiographical memory knowledge base*. These two components can function independently of each other; however, the system is formed and autobiographical memory occurs only when the working self and the autobiographical memory knowledge base are conjoined and work together (Bower, Black, & Turner, 1979).

An essential idea of their model is that autobiographical memories are transitory dynamic mental constructions generated from an underlying knowledge base. The autobiographical memory knowledge base contains two types of information: autobiographical knowledge and abstract levels of abstraction (Conway & Pleydell-Pearce, 2000, p. 271); the latter is "the summary records of sensory-perceptual-affective processing" (Conway, 2005; Williams et al., 2008, p. 37). These two can be regarded as two forms of mental representations, which are respectively responsible for, and can be made cognitive and phenomenally accessible to, the processing of the semantic and the episodic memory systems (recall the contribution of the interplay of episodic and semantic memory to autobiographical memory in §2.1.5).

One feature of autobiographical memories is that they contain knowledge at different levels of specificity. According to the SMS (Conway, 2005; Conway & Pleydell-Pearce, 2000), autobiographical knowledge is organized in hierarchical knowledge structures[15] ranging from highly abstract conceptual knowledge (i.e., themes and lifetime periods) to event-specific and experientially close conceptual knowledge). On the most abstract level is a structure termed "life story" which contains general factual and evaluative knowledge about the individual. One's life story contains the general factual and evaluative knowledge about oneself, which is organized by different themes referring to different self-images. For instance, my life story may consist of themes about my self-images as a daughter, a student and so on. These different self-images are linked to different parts of the autobiographical memory knowledge base. That is, my self-image as a daughter is connected to my memories of my parents when I was a child. There are several lifetime periods linked to each of the themes. In this way, the theme of being a student may be linked to the period of elementary school, to high school, and to university. Each of the periods also connects with general events, such as doing homework, sitting examinations, and conducting experiments.

In addition to the autobiographical memory knowledge base, the other main component of SMS is the "working self". The main function of the working self is goal management (e.g., coordinating goal processing, maintaining goal compatibility, and goal prioritization) and maintaining the coherence and correspondence of memory. It is a set of currently active goals or self-images that are organized, through working memory, into a goal hierarchy, which is "a highly

[15] The way themes, lifetime periods, general events, and episodic memories are organized can be found in Figure 5 in Conway (2005).

complex goal-sub-goal hierarchy of interlocked negative and positive feedback loops in which goals are represented at different levels of specificity" (Conway, 2005, p. 596). The purpose of the goal hierarchy is to regulate behavior by reducing discrepancies between desired goal states. It is through the goal hierarchy and the working self that new information can be organized into long-term memory and stored information can be accessed and constructed as memory (Williams et al., 2008, p. 37).

To return to the issue of this section: What are the functional constraints for accounting for the emergence of autobiographical memory and an ASM? What constructs our autobiographical memory and ASMs and affects the way they appear? According to the SMS, autobiographical memory emerges under the competing constraints of coherence and correspondence. According to Conway et al. (2004), "[a] central principle of the SMS framework is that memory is a product of the tradeoff between the separate but competing demands of coherence and correspondence". Coherence is the more important demand, and in an extreme environment in which these constraints cannot be mutually satisfied, the system tends to sacrifice correspondence in order to secure coherence.

Coherence and global veridicality result from the interplay of several factors. First, what is directly related to is the organization of the mental representations in long-term memory or, in Conway's terms, in the autobiographical memory knowledge base. As we have seen in §2.3.1, memory is not stored in the "Store-house" (Locke, 2008) and reproduced during retrieval. Instead, it is stored as elements and reconstructed later (Schacter et al., 1998). The organization of these elements determines how easy it is for a memory or mental simulation to occur if it supports current contexts and goals. Alzheimer's patients are an example of disorganized mental representations. The loss of such organization results in an apparently random combination of elements of past experience, which leads to difficulty in recombining those elements in a coherent and globally veridical fashion.

However, in contrast to Conway's account, I differentiate coherence constraint into synchronic and diachronic coherence and suggest that *synchronic coherence*, *diachronic coherence*, and *global veridicality* are three functional constraints for an ASM. I will introduce these one by one.

First, synchronic coherence (i.e., coherence for Conway) can be found in biases of memories. For example, supporters of a football team tend to recall the

matches that confirm their belief that their team is highly skilled and has the best team cooperation. According to Conway (2005), coherence is

> […] a strong force in human memory that acts at encoding, post-encoding remembering, and re-encoding, to shape both the accessibility of memories and accessibility of their content. This is done in such a way as to make memory consistent with an individual's current goals, self-images, and self-beliefs. Thus, memory and central aspects of the self form a coherent system in which, in the healthy individual, beliefs about the knowledge of, the self are confirmed and supported by memories of specific experiences. (p. 595)

Synchronic coherence is maintained by modulating the construction of memory. Beike and Landoll (2000) report on their observations of strategies adopted for maintaining coherence, such as outweighing, justification, and closure. When an individual recalls a happy memory during what is otherwise considered an unhappy period, the happy memory may be regarded as exceptional (outweighing). The individual may also consider the remembered event to be a justifiably unhappy event in an otherwise happy period, or she may just gain some closure on the dissonant memory and stop further processing.

Extreme violation of coherence only arises in cases of psychological and psychiatric illness. Delusions and confabulations resulting from psychological disorders suggest the importance of the synchronic coherence of autobiographical memory. It can be seen from cases of DID that a certain degree of synchronic coherence is required in order to maintain an ASM. In order to maximize the overall synchronic coherence, DID patients have created other ASMs that are functionally separated from each other but are themselves coherent (Metzinger, 2004, pp. 522-528). This results in multiple personalities from a third-person perspective.

Second, in addition to the synchronic coherence that Conway proposes, there is another functional constraint—diachronic coherence. This refers to consistency between the contents of ASMs constructed at different times. The dissociation of synchronic and diachronic coherence can be found in patients who suffer from dementia. When these patients confabulate, they may at different times create dramatically different ASMs that result in distinct mental autobiographies. Nevertheless, at each point in time, the patients strive to maintain the synchronic coherence of that ASM.

Why is there diachronic incoherence in an ASM? Incoherence can result from (1) the way information is encoded, or (2) distortion of information when undergoing reconsolidation. First, it is worth noting that the content of one's ASM is not a record of information about oneself; instead, it is information about how the world and oneself are represented. How we conceive the world and ourselves depends on one's current content of PSM—external information (e.g., perception) and internal information (e.g., autobiographical representations from one's ASM) are integrated. Therefore, one may conceive the same thing differently at different times. Second, except for the biases resulting from the first encoding, information can be distorted every time it is retrieved and reconsolidated. We often manipulate mental representations to increase the overall coherence of the content of a currently active PSM, and incoherence may consequently emerge.

Accordingly, we do not have a diachronically coherent ASM. Diachronic incoherence can emerge to different degrees. Compared to dementia patients, whose ASMs are acutely diachronically incoherent, most of us maintain the ASMs that are comparatively coherent with previous ASMs. That is, diachronic incoherence does not occur solely in patients with mental disorders; rather, slight diachronic incoherence can exist from time to time for everyone. For instance, we change our character during our lifetime, and may deceive ourselves by reinterpreting past events. Furthermore, diachronic incoherence is likely to happen when there is a change in (social) environment. A person might show different characters based on different environmental or social needs.

Compared to synchronic incoherence, which can be detected phenomenally because it directly leads to suffering, diachronic coherence is a constraint that is less easy to detect subjectively. That is, as long as there is a synchronic coherent ASM, one owns knowledge about oneself—one's past, future, preference, values and other characters—through the ASM, and will honestly believe that one's past mental autobiography is informed by the current ASM. Unless there exists external clues (e.g., records, traces left from the past, or other agents) that point to an inconsistency, diachronic coherence or incoherence is unlikely to be discovered by the subject.

In spite of the fact that diachronic incoherence is not as urgent as synchronic incoherence, which the system automatically strives to avoid, diachronic coherence is essential for being a person. As we will see below, the concept of personhood refers to the property of the whole system, which emerges when the requirements for membership of a moral community are fulfilled (see §3.3.4). While according to

the normative concept of personhood, personhood relies on a mutual conception of one another as agents; a degree of diachronic coherence is certainly necessary. This necessity is evident in how we regard the shift of ASMs in DID. We intuitively regard them as different persons realized by the same organism.

The third constraint is global veridicality. This refers to the degree to which the important functional feature of the ASM corresponds to the truth. I disagree with Conway's (2005) idea of "correspondence": According to Conway, an ASM with the satisfaction of the constraint of correspondence is important for an organism, because, from an evolutionary perspective, a memory system that does not maintain an accurate record of goal processing is unlikely to survive (p. 596). Memory must then correspond to a real past experience. However, this is not entirely true. As has been shown in §2.3.2, memory has evolved to be more flexible, regardless of the cost of losing global veridicality. The advantage of such flexibility is that it allows the memory to adapt to changing contexts and goals. A high degree of correspondence or veridicality is therefore not always beneficial to the subject (e.g., self-enhancement as a kind of self-deception).

Therefore, I suggest the alternative concept of global veridicality. It is global instead of local because it does not require the system to construct a content that is identical to the content of past representation, but only the similarity in the general structural features and important and functional features in the sense that it is critically relevant to the present context and current goals. Such an idea is expressed by Martin and Deutscher (1966), who argue that, "the state or set of states produced by the past experience must constitute a structural analogue of the thing remembered, to the extent to which he can accurately represent the thing" (p. 191). They used Wittgenstein's example of the structural analogy between music and the groove in a gramophone record in order to depict the relation between memory content and what was experienced. For them, these do not need to be perfectly structurally analogous:

> Perhaps there is no sense to the idea of mirroring all the features of a thing, for there may be no sense in the notion of all the features of anything. But it is enough for our purposes that we can make sense of the idea of an analogue which contains at least as many features as there are details which a given person can relate about something he has experienced. (p. 190)

That is, I suggest that to which degree a global veridicality is required depends on the current use of the representation. The structural feature may be commonly required, whereas the local features may be only be required if the current goal is related to a detailed investigation of a past event.

To determine the veridicality of a memory, we need to examine the content of memory simulation and the object of the simulation, that is, the content of the past representation. However, how do we know whether the content of our memory is veridical, that is, whether it truly corresponds to a past episode? For instance, I seem to remember something I experienced alone, but without any evidence or anyone to support the veridicality of my experience, doubts are warranted concerning whether this actually happened or whether it occurred only in my imagination.

> The very idea of truth in memory, and the attendant possibility of error, implies that we are naturally realists about the past: but this fact about us doesn't dictate answers to questions about just how, or how often, we do remember the past truly. (Sutton, 2010)

This question relates to the issue of how memory can be distinguished from the imagination. As with the skeptical objection raised against representationalist theories of memory, we have no way of knowing exactly what has happened in the past. This is especially problematic for episodic memory, because for an episodic simulation to be veridical, one is required to investigate the content of past representation, that is, what the subject experienced in the past. Some experiences can be recorded in certain ways—such as in diaries—but some phenomenological aspects of experience (e.g., its what-it-is-likeness) cannot be fully preserved in this way.

Failure to satisfy the constraints of diachronic coherence and global veridicality often occurs when one is suddenly placed in a radically distinct environment, when one may feel sort of “disorientated”. This happens because the organization of the mental representations in one’s long-term memory is shaped in a way that accommodates the context one is most often engaged in. Nevertheless, thanks to the constructive nature of memory and re-consolidation, one can flexibly change the organization to fit current needs. Therefore, contexts and goals can partially determine how easy it is to form a diachronically coherent and globally veridical memory.

In addition, most of us can, under normal circumstances, be corrected by our environment. For instance, if I remember having cold melon soup in a certain restaurant, but all the evidence shows that this restaurant has never served any kind of soup, I will find and correct the mistake in my retrieved memory. Such mechanisms allow us to increase the global veridicality (and diachronic coherence) of constructed memory. However, for some patients, such mechanisms do not exist. For instance, Breen et al. (2000) report the case of RZ, who suffers from reverse intermetamorphosis and holds the delusional belief that she is her father, and occasionally her grandfather. It seems impossible for her to correct her belief, despite all evidence suggesting that she is in fact a 40-year old woman.

3.3 What I Am

What am I? This section aims to introduce two concepts—"person" and "human animal"—by reviewing two kinds of answers to the question posed above. There are two kinds of approaches to answering this: The psychological approach aims to answer it through specifying the essential psychological characteristics, e.g., personhood; the biological approach, on the other hand, resorts to biological criteria. The former may answer that we are persons, while the latter will answer that we are human animals. It is natural that neither would deny that we are persons as well as animals; [16] therefore, the critical discrepancy between them is what we fundamentally are or, in another words, what determines my existence. For those who embrace a psychological approach and contend that I am fundamentally a person in virtue of having psychological properties, my existence condition is derived from my fundamental nature of being a person: It is impossible that I could exist without being a person. In the following sections, we look at the arguments for these approaches.

3.3.1 Being a Human Animal: The Biological Approach

The biological (or somatic) approach is also called animalism (DeGrazia, 2005b; Olson, 1997; Snowdon, 1991). According to animalism, we are essentially animal. Animalists leave the question of personhood open: One can endorse animalism and at the same time hold that we are persons: Being an animal organism or being a

[16] There is a discussion about whether one can be both person and human animal. Philosophers such as Lynne Rudder Baker have differentiated different usages of "is": The term "is" can be used as the relation of identity or its constitution.

person are regarded as two different states of being, but we are fundamentally the former.

Olson (2003a), one of the main proponents of animalism, presents the following argument for animalism:

1. $(\exists x)$(x is a human animal & x is sitting in your chair)
2. (x)((x is a human animal & x is sitting in your chair) $\supset$ x is thinking)
3. (x)((x is thinking & x is sitting in your chair) $\supset$ x = you)
4. $(\exists x)$(x is a human animal & x = you) (pp. 325-326)

There is a human animal that is sitting in your chair (1). Every animal that is sitting in your chair is thinking (2). And that which is thinking and sitting on the chair is you (3). Therefore, you are a human animal (4). Based on this argument, the problem of personalism that Olson argues for is that if you are a human animal, one cannot also fundamentally be a person which is not identical to a human animal, otherwise there will be two entities—a human animal and a person—sitting on your chair, which is obviously not true.

The objections to animalism, concerned by Olson (2003a) include the following: One could argue that there is no human animal at all (against premise 1) or human animals cannot think (against premise 2). Idealists, or those who advocate the view that nothing can have different parts at different times (e.g., Chisholm, 1976, pp. 145–158), deny the metaphysical existence of organisms and accordingly will refute the first premise. As for premise 2, some argue that animals cannot think (e.g., Shoemaker, 1984, p. 92–97); however, "thinking" doesn't seem to be a critical element in this argument: According to animalism, thinking has nothing to do with being a human animal or with being you. One could substitute "x's heart is beating" for "x is thinking" or replace thinking with having proprioception, breathing, or reading.

However, although this argument is sound, it's not informative: It does not argue for animalism, which claims that our metaphysical nature is being animals. This argument does not conclude that we are essentially animals, but only that there exists a human animal and that it is you. Olson is right that at the very moment when I'm thinking about the argument, I'm an animal; however, am I still an animal when I am in a dreamless sleep? In general, it says nothing about a patient in persistent vegetative state (PVS), who cannot think about this argument, and therefore can never be the "you" in the argument.

Furthermore, if Olson claims that such an argument can be used to successfully argue for animalism, let us consider my modified argument below:

1. $(\exists x)(x$ is a citizen & x is sitting in your chair)
2. $(x)((x$ is a citizen & x is sitting in your chair) $\supset x$ is thinking)
3. $(x)((x$ is thinking & x is sitting in your chair) $\supset x =$ you)
4. $(\exists x)(x$ is a citizen & $x =$ you)

Does it result in "citizenism"—that one's essential nature is being a citizen? It merely concludes that I am a citizen, which is true, rather than resulting in my metaphysical nature of being a citizen. "Citizen" can also be replaced by "person" and the argument concludes that we are persons, but it does not lead to personalism.

In fact, few people will deny that we are animals. The kind of relationship between being a human animal and being a person other than identity has been provided. For instance, Glannon (1998) suggests

> a person is neither identical with any one of these stages nor with the organism itself. The embryo or presentient fetus is a potential person, not in the sense that it becomes a person, which implies numerical identity, but only in the sense that it has the potential to develop the biological structures and functions necessary to generate the consciousness and mental life constitutive of personhood. (p. 190)

Furthermore, Lynne Rudder Baker (2002), as an opponent of animalism who argues for the ontological significance of persons, agrees that we are animals. In the next section, we will have a detailed look at her constitution view and her idea of the relation between persons and animals.

Because of the metaphysical claims of animalists, when speaking of the issue of transtemporal identity, they either refer to the issue as "human identity" (DeGrazia, 2005b) or use the term "personal identity" for the sake of specifying the exact question in which philosophers have been interested (Olson, 1997, pp. 26-27). For them, there is no such thing as "personal" identity: Olson (1997), for instance, denies that there can be a general criterion of identity for all persons. We will discuss the issue of animal identity together with personal identity in §3.4.

3.3.2 Being a Person: The Psychological Approach

Different from animalism, the psychological approach claims that we are essentially a being with certain psychological characteristics, which is usually referred to as "personhood". What does the concept of person or personhood mean? We regard ourselves as persons. We have intuitions of what a person is, since we can identify some members in this set. However, when it comes to patients with significant cognitive deficits, children under development, or the humanoid robot in *Ex Machina*, this intuition becomes elusive.

One of the most notable proponents of psychological approach, John Locke (2008), made the distinction between "person" and "man" (or "human animal"):

> We must consider what idea the word it is applied to stands for: it being one thing to be the same substance, another the same man, and a third the same person, if person, man, and substance, are three names standing for three different ideas. (p. 37)

According to the Lockean view, the concept of person is not coextensive from the concept of human animal. A man is an animal of a certain form (p. 38); a person, on the other hand, exists as "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places" (p. 39).

According to such a distinction, there are humans that are not persons. As persons emerge later than human organisms (Glannon, 1998, p. 190), infants and children under a certain age do not have full personhood. They are human beings: They belong to the species *Homo sapiens*. We have the intuition that these non-person human beings have some rights (e.g., rights to live or rights to be free from suffering) but are free from some moral charges or legal rights and responsibilities (e.g., the right to vote). Patients in PVS may also be considered lacking personhood. On the other hand, whether there could be non-human persons—persons that do not belong to the species—is contentious. Would and should artificial intelligence capable of thinking and acting just like us be considered persons? This question relies on what makes a being a person. We will look at the moral significance of being a person in the next section, and focus on the question of criteria for personhood in §3.3.4.

Following the distinction between person and man, Locke also distinguishes person from self:[17] "Person" is a forensic term of "self". "Person" only belongs to intelligent agents capable of a law, happiness, and misery (Locke, 2008, pp. 50-51), whereas "self" is defined as

> […] that conscious thinking thing, whatever substance made up of (whether spiritual or material, simple or compounded, it matters not), which is sensible or conscious of pleasure and pain, capable of happiness or misery, and so is concerned for itself, as far as that consciousness extends. (p. 45)

This suggests that "person" is the normative or legal aspect of self, and "self" refers to a form of consciousness that carries out certain mental capacities. Such mental capacities allow one to "be qualified" as a person and to be subject to rights and responsibilities.

Lynne Ruder Baker (2000, 2013), following Locke's distinction between person and animal, develops the constitution view. Baker claims that we are both persons and animals, but we are fundamentally persons. Recall Olson's concerns in the last section: Does Baker's idea implies that there are two entities occupying the chair, namely a person and an animal? According to her view, one is a person in virtue of having a *first-person perspective*, and one is a human person in virtue of being a person constituted by a human animal, which is material.

First, Baker distinguishes weak first-person phenomena from strong first-person phenomena (2000, pp. 60-69): These are exhibited when the organism is respectively *making* a first-person reference (weak) and *attributing* a first-person reference to itself (strong). Weak first-person phenomena are exhibited by sentient beings whose behavior can be explained by perspectival attitudes. These organisms act from their own perspectives. For instance, a non-human animal experiences in a way that it feels itself at the center of the experience. This is the weak sense of first-person perspective.

According to Baker, organisms which merely exhibit weak first-person phenomena are not self-conscious but only conscious: They have a rudimentary first-person perspective, rather than a robust one (2013, p. 40). To be fully conscious, one requires a first-person perspective, which results in a robust first-person perspective: One has to be able not only to recognize oneself from a first-

[17] Barresi and Martin (2011), using "self" and "person" interchangeably, think that Locke doesn't differentiate between these two concepts.

person point of view, but also to think of oneself as oneself (2000, p. 64). That is, the subject can "conceptualize the distinction between himself (and everything else) from a third-person point of view and himself from a first-person point of view" (p. 67). Organisms with robust first-person perspectives are then capable of reflecting on their own thoughts and desires.

Baker (2000) explains the differentiation of strong and weak first-person phenomena in terms of a grammatical distinction between "I" and "I*": "I" is a first-person pronoun, which is used to directly refer to oneself; "I*" is first-person indirect reflexive pronoun, which is used to refer to oneself characterized in the first person, by means of a self-concept—a concept of oneself from one's point of view (2013, p. 36). The weak first-person phenomena refer to the way that organisms act from a centered perspective, which can, for instance, be expressed as "I am hungry"; as for strong first-person phenomena, they require that the subject be able not only to have a first-person thought, but to attribute first-person thoughts to herself, for example, "I think I* am hungry" (2000, p. 67). Baker's "I*" is extended from "he*"—quasi-indicators invented by Hector-Neri Castañeda (2001). Like quasi-indicators, "I*" is embedded in a "that"-clause, following a psychological verb. I*-thoughts are thoughts that can be expressed by I*-sentences, which are the sentences containing "I*". To have an I*-thought is to have the ability to attribute the first-person reference indirectly to oneself.

Second, on the constitutive view (Baker, 2000), the definitive property of a person is the capacity to have a robust first-person perspective. For one to have the capacity for a robust first-person perspective means that one has the structural property required for a robust first-person perspective, and one has either exhibited a robust first-person perspective before or is in an environment conducive to the development and maintenance of such a perspective (p. 92). A person doesn't have to be a human: A human person is a person constituted by a human animal.

As such, third, the relation of constitution is not identity but "contingent identity": "Identity is necessary; constitution is contingent" (Baker, 2002, p. 372). Unlike the relation of identity, constitution is an asymmetric relation: The human animal constitutes the person; the person does not constitute the human animal. According to the constitution view, our *primary kind*—what we most fundamentally are and our persistence condition—is a person; the *primary-kind property* is the property in virtue of which we exist. We could not exist without our primary-kind

property, which is the capacity for a robust first-person perspective.[18]

In summary, the constitutive view holds that we are fundamentally human persons: We are persons in virtue of possessing the capacity for a robust first-person perspective; we are human animals in the sense that we are constituted by human animals. To have a robust first-person perspective is to be able to attribute a first-person perspective to oneself—to think of oneself as oneself.

The distinction between weak and strong first-person phenomena is also depicted by the SMT (Metzinger, 2004; see also §3.1). For one to have a rudimentary first-person perspective, the given system has to form a PSM. For example, what can be illustrated in language—"I am hungry"—is a conscious self-representing system with the integrated content of being hungry. As in the robust first-person perspective, the special relation of oneself and the mental event is also represented. Take "I think that I* am hungry": One being the owner of the thought is also integrated into the conscious self-model, which now turns into a PMIR.[19]

3.3.3 The Moral Significance of Personhood and Selfhood

We are essentially a PSM, and so are some other animals.[20] What distinguishes us from those animals is the special kind of self-model that has been developed within us. This enables us to *become persons* and to have a different moral status from other creatures. As discussed in the last chapter, Baker has pointed out a critical criterion: the robust first-person perspective. This section illustrates the different moral significances of selfhood and personhood—two critical states of the system.

The first concept that is morally important is simply being a PSM, that is, a being that is capable of having phenomenal states. As I will discuss later in the next chapter (§4), one of the moral demarcations is suffering: According to negative utilitarianism, we ought to minimize overall suffering. Only systems which are transparent, self-representing (i.e., PSMs) are capable of suffering or any other form of negative experience equivalent to human suffering. As such, the set of systems satisfying these representational and functional constraints is morally distinct from others in that when considering the moral issues (e.g., the moral permissibility of

[18] For a complete definition of "constitution" and a detailed explanation of "primary-kind", see Baker, 2000, 2002.
[19] For a detailed analysis of I*-thought with SMT, see Metzinger (2004).
[20] It is not yet clear what kind of animals possess not only a self-model but also a PSM. Some (e.g., mammals) obviously have a simpler PSM than that of most human beings, but for those (e.g., fishes) whose central nervous systems are widely divergent from ours, we are not sure if they are capable of having a PSM.

CE), the members of this set (PSMs)—as well as potential members (potential PSMs)—are under consideration. Accordingly, this leads to some moral and legal consequences. For instance, some (e.g., elephants) have rights not to suffer, while others (e.g., rocks) do not have such rights. (For more detailed discussion of the moral significance of self- and identity-related concepts, see §8.)

Personhood is another important property that makes a system not only qualitatively different but also morally distinct from a system which merely satisfies the functional constraints to be considered a PSM. Though what criteria belong to the set of necessary and sufficient conditions for being a person is still debatable (e.g., intelligence, rationality, self-awareness, memory, future-thinking, linguistic ability, and sympathy), one essential and basic criterion is the capacity for subjectivity, according to SMT (Metzinger, 2004), or a robust first-person perspective in the constitutive view (Baker, 2000). Harry Frankfurt (1982), for instance, proposes that what makes persons differ from other non-person creatures is the structure of will (p. 6): Persons are capable of generating "second-order volitions" in addition to "first-order desires" (see §7.3.5).

Personhood opens up a moral dimension. It precedes moral responsibilities and additional rights. Non-persons, such as a coffee machines, will never be praised or blamed (or punished in legal aspects) for good- or ill-performance, because the sense of responsibility in the coffee machine is but a causal connection (Eshleman, 2009). On the other hand, persons will be praised or blamed for their performance, for we are responsible for doing or not doing something, in the sense that in certain circumstances there are roles that we are obligated to fulfill. In short, persons are subject to moral responsibilities. Accordingly, the concept of a person plays a significant role in related bioethical issues, such as abortion, euthanasia, and stem cell research.

Why is the concept of person crucial? What is the moral significance of being a person? Persons—organisms equipped with robust first-person perspectives (Baker, 2000) or phenomenal subjectivity (Metzinger, 2004)—are capable of regarding themselves as themselves, and of reflecting upon their own thoughts. This is the necessary criterion for opening up the moral dimension. First, they are capable of reflection. Second, they not only recognize themselves as a thinking being, they also recognize the existence of other agents. With empathy, they have the capacity to simulate the mental and phenomenal states of other organisms. Third, episodic memory provides the organism with a higher degree of flexibility (§2.3.2): The organism is not merely capable of performing stimulus-response behaviors, but

reason and action are also available. Last, a self-world boundary allows them to have the concepts of "true" and "false", and interaction with other agents results in the emergence of normative concepts.

3.3.4 A Normative Concept of Personhood

Following Locke's idea of treating "person" as a forensic term, "personhood" is a property of the whole system when the system fulfills the requirement of membership of its moral community and shares moral responsibilities and the rights that accompany those responsibilities. Here I propose a normative concept of personhood. Like *selfhood*, it is the property of the whole representational system, which can be accounted for by the satisfaction of a set of particular functional constraints: For a system to become a phenomenal self, it has to be a transparent self-representing system; for a system to become a person, not only are the constraints previously mentioned involved but also other additional constraints, for instance, those for being considered as a PMIR. However, unlike the concept of selfhood, how can a system become a person—in another words, what are the criteria that the system has to achieve to turn into a person—this depends heavily on how the concept of person is conceived.

Why is being a person important? There are two reasons why being a person is important: One relates to the third-person; the other to the first-person. The third-person conception of a person concerns whether a being is a person through third-person assessment of its capacity, through its objective properties, such as moral behavior, emotional response, and rational choice. It is based on a capacity-theoretic conception of responsibility, according to which one must exhibit relevant mental capacities in order to be responsible for one's actions (Glannon, 2007, p. 57). In general, this is how we decide if one belongs to the "moral community", in which members are subject to praise and blame or agreed moral responsibilities and rights. For instance, in most cultures children under a certain age (e.g., 16, 18, or 20 years old) do not fully participate in the moral community and are not granted some responsibilities (e.g., criminal responsibility) and rights (e.g., suffrage). We also consider patients with particular psychological disorders or those under altered mental states in this way. This is mainly based on our third-person judgment that the subject in question does not have the capacity of personhood either temporarily or permanently.

On the other hand, we also consider the first-person viewpoint when we think about whether we are a person: Let's call it *the sense of personhood*. For most

of us and under most circumstances, we have a sense of being a person, and the question, "Am I still a person?" doesn't occur to us. Nevertheless, doubt emerges when one is losing one's sense of personhood. For instance, the doubt of a dementia patient is illustrated from the first-person point of view in *Out of Mind*, a novel by J. Bernlef (1989):

> All of a sudden I had to translate everything into English first, before I could say it. Only the forms of sentences came out, fragments, the contents had completely slipped away. Furiously I glare into the room. I seem to lose words like another person loses blood. And then suddenly I feel terribly frightened again […] I quickly lie down on the settee and close my eyes. A kind of seasickness in my mind it seems. Under this life stirs another life in which all times, names and places whirl about topsy-turvy and in which *I no longer exist as a person*. (p. 63; my emphasis)

The feeling of personhood fades when one finds oneself lacking certain mental capacities (e.g., reasoning, self-control, or memory) and therefore has difficulty dealing with one's ordinary life tasks. One might consider oneself unqualified in the moral community, and no longer belonging to the category of "person". The distinction between self-directed ascriptions and other-directed ascriptions of responsibility (Eshleman, 2009) is respectively related to a sense of personhood and the concept of a person.

To return to the issue of the concept of person, the criterion of what makes a person (third-person conception) co-evolves with a sense of personhood (first-person conception): that is, how we acknowledge others as persons influences how we acknowledge ourselves as a person and *vice versa*. Our shared concept of person—how we recognize each other as persons—originates from our sense of personhood. We regard those with similar mental and moral traits as being of the same kind as us. On the other hand, the concept of person influences our attribution of praise and blame (or responsibilities and rights) as well as policy-making. This affects our judgment of ourselves: The sense of personhood indicates our idea of ourselves as capable of being a member of the group of persons. Moreover, our shifting environment also influences our concept of personhood, as well as our sense of personhood. For instance, one might expect in the future to make the same judgment through the capacity to integrate technology into one's cognitive life. For instance, if the brain stimulation technology is further improved and easily

accessible for everyone to integrate it into her daily life, the interaction between the agents in this community will require a different level of cognitive capacity. Cyborg is another possibility. If these forms of manipulation of human cognition are in widespread use, those who refuse to or are not capable of using cutting-edge innovation and are not updating fast enough may be excluded from a newly-developed criterion of personhood. This leads to an interesting point: To be able to keep up with and to be open to new manipulation on human brain and cognition may become more central and may one day become necessary to stay in the moral community and to be considered a person.

In other words, the criteria for personhood can alter with a change in the states of the system. According to Damasio, the nonconscious neural signaling of an individual organism results in the emergence of a proto-self, which permits the core self, and which in turn allows for the autobiographical self. He further suggests that "conscience" is another emergence that brings a qualitatively different self (1999, pp. 230-231, Figure 10.1). Based on the SMT (Metzinger, 2004), different forms of self-models and their relation with the concept of personhood are shown in Figure 3. With the satisfaction of different functional constraints, the system is capable of forming a more complicated form of self-model, and with the emergence of new forms of self-models the criteria required for interaction with other more complicated self-models change.

Therefore, different from the concept of selfhood, the concept of personhood can only be reduced *in an indirect way*. It is described by Metzinger (2004) as follows:

> The concept of a "person," however, does not simply refer to some complex, but objective representational property. Personhood cannot be naturalized in a simple and straight-forward way, because the concept of a person contains domain-specific and semantically vague normative elements. Why is this so? Persons never are something we find *out there*, as parts of an objective order. Persons are constituted in societies. If conscious self-modeling systems *acknowledge* each other as persons, then they are persons. (p. 601)

Consequently, the mental capacities required for being a person are different from culture to culture and may change through time.

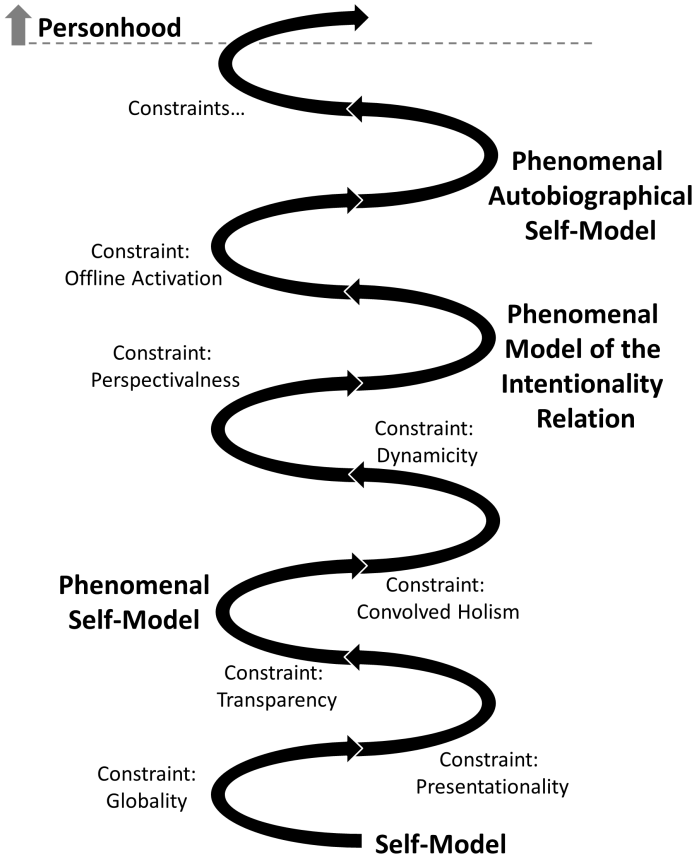

Figure 3. Personhood and the self-models.

Thus, to understand the criteria of being a person is a matter of empirical investigation. One is required to be involved in a moral community to look into what is required to be treated as a moral agent. Nevertheless, Dennett (1976, pp. 177-178) has suggested six conditions which are necessary for moral personhood: (1) being a rational being, (2) beings to which states of consciousness are attributed, (3) being treated with certain attitudes, (4) being capable of treating others as persons, (5) being capable of verbal communication, and (6) being self-conscious (in the way that no other animals are capable of). These conditions can be regarded as the necessary conditions required for a group of agents to form a moral community. The current society demand more cognitive capacity for joining the membership.

It is noteworthy that the distinction between human animal and person triggers a reconsideration of the concept of death. Glannon (2007) proposes three criteria of death: the higher brain, whole-brain, and brain stem criteria (p. 148-156). The first concept of death defined by the high brain criterion refers to the end of all higher brain functions. This is the concept pertains to persons, instead of human organisms. The other two concepts of death, pertaining to human organisms, respectively refer to the end of critical functional processes regulated by the brainstem and the end of integrated somatic processes. Glannon defines persons as beings with the capacity for consciousness. Thus, the higher brain functions refer to the functions required for a system to generate a PSM. However, based on my normative concept of personhood, the criteria which pertain to persons are determined by the current moral community and can be different from culture to culture and alter in the future.

Martha J. Farah and Andrea S. Heberlein (2007), based on their findings, propose that the concept of personhood does not correspond to any real category of objects in the world; it is the product of an evolved brain system that develops innately and projects itself automatically and irrepressibly onto the world. Scientists can find an objective basis for "plants", for instance, but not for "persons": Such a difference results from the lack of "objective reality" of the category of person (p. 44).

Farah and Heberlein (2007) compare the concept to "phlogiston": In 17th and 18th century, some phlogiston theorists believed that phlogiston was contained in combustible substances, and burning was the process by which phlogiston left those substances. Now we know that there is no such thing as phlogiston, and burning is a kind of oxidization. The concept of "phlogiston" resulted from the correspondence between phlogiston theorists' representations of phlogiston, and categories of events in the world that corresponded in a systematic way. The example shows that "mental representation can exist and be activated by stimuli in systematic ways without picking out fundamental categories of the natural world" (p. 44). Thus, the concept of a person is different from concepts such as "*Homo sapiens*" that can be picked out from the world according to some facts. Instead, it is similar to the concepts of "patient" or "citizen" that require a normative claim.

According to Farah and Heberlein (2007), science is unable tell us what a person is, but is capable of explaining our intuition of the distinction between person and non-person. First, we have a set of specialized brain systems for representing persons, such as the visual recognition of human face (fusiform gyrus),

the human body (an area on the fusiform gyrus but distinct from face recognition area and the other on the lateral surface of the brain near the temporoparietal juncture), bodily movement (another part of temporoparietal junction), and representing other's mental state (medial prefrontal cortex) (pp. 40-42). These regions are referred to as "the social brain" and are considered by the authors "a network for person representation". Second, the person network has a high level of automaticity and innateness; that is, this system is triggered by certain stimulus features automatically and is genetically programmed.

Accordingly, personhood is a kind of illusion created by our brains. This concept is dynamic: It alters with changes in our conventions, norms, and preferences and may vary from culture to culture. There is a close connection here with the way we understand the concept of "normalization", which has a dynamic interaction with "normality" (Metzinger & Hildt, 2011, pp. 247-248; see §4.3.1). This will be discussed further in the next chapter. The main idea is that the concept of a person is not fixed and may change with the alteration of our minds, which is a noteworthy part of CE discussion.

Because of the dynamicity of "personhood", CE at a larger scale in a community can influence (e.g., speed up) the evolution of personhood. Walter Glannon (2011) suggests that massively strengthening the capacity to respond to reasons may lead to a higher level of norms of moral sensitivity, and alter the "reasonable person standard" in criminal law (p. 120). However, this influence in terms of personhood can only occur if a large number of the moral community undergo CE (e.g., due to the high accessibility of cognitive enhancers), and the affected cognition is directly (and indirectly) relevant to their moral capacity.

3.4 Transtemporal Identity

The issue of *transtemporal identity* concerns the identity relation across time. This issue is usually referred to with the term "personal identity". This is likely to lead to confusion, and it seems to presuppose that one has already endorsed the view that one is essentially a person, and that transtemporal personal identity is what determines one's existence across time. Animalists will object to such an assumption: Eric T. Olson (1997), for instance, uses the term "personal identity" to refer to the issues that have been discussed by philosophers, but what he actually has in mind is "animal identity" (pp. 26–27). Likewise, David DeGrazia (2005b), refrains from using this term, and instead talks about "human identity". Here I use

the term “transtemporal identity” to refer more generally to such relation across time.

Transtemporal identity of what? I contend that we have different transtemporal identity relations. *Transtemporal personal identity* concerns the question of how one person at one time is identical to a person at another time. *Transtemporal human identity* considers how one human animal at one time is identical to a human animal at the other time. Beside this, there are also other transtemporal identities, such as citizen identity and student identity. In addition to the criteria of each identity relation, the issue is, therefore, the different roles of these different kinds of transtemporal identity: How are they differently important to us? What is their moral significance? This will be the basis of our discussion in §8.

3.4.1 The Concept of Identity

What is identity? Some have mistakenly considered identity as a property or sometimes personality:

> There is a common misunderstanding about the notion of identity: Identity is not a thing or a property which you can have, like a bicycle or the color of your eyes, but a relation. Identity is the relation in which an entity stands to itself. (Metzinger & Hildt, 2011, p. 254)

To say that things share a relation of identity or that things are identical means that they are the same. But there is a distinction between qualitative and numerical identity: Things with *qualitative identity* share properties and are exactly similar to each other. On the other hand, *numerical identity* is to say that that there is only one thing, and thus a thing can only be numerical identical to itself. The question of numerical identity can be understood as how many times we count each thing: If we point twice, first at person X and then at person Y, X and Y are numerically identical if and only if X and Y are pointed at twice. X and Y are not numerical identical if and only if X and Y are pointed at once.

Concerning the criterion of identity, what is the standard by which identity is to be judged? Criteria can be considered synchronically or diachronically. The former consider the conditions under which coexistent objects are identical, whereas the latter questions how an object remains the same over time. Moreover, transworld identity is the notion of identity across possible worlds. Here we will mainly focus on trantemporal identity.

3.4.2 The Sense of Identity

It is also worth noting that there is a distinction between a *sense of identity* (or sense of continuity) and metaphysical relation of identity. The last section concerns the latter, and we will investigate it further in the following sections. The former, a sense of identity, unlike the metaphysical relation, is a phenomenological concept. It refers to our feeling that we are the same as a particular being in the past. We experience ourselves not only as existing at the present moment but also in the past and potentially in the future.

The sense of identity refers to the subjective feeling of being someone with some characteristics or the phenomenal experience that I exist diachronically in the past as well as the future. The defining criteria for the sense of identity at the phenomenological level include (1) conceiving oneself as a subject with relations to objects, events, or other agents; (2) being partially mentally disconnected from the current environment; (3) being situated at a mentally constructed spatiotemporal location of "now" and "here" with a past and a future; and (4) having a sense of ownership of past experiences or future projections. The loss or the deviant forms of the sense of identity can be found in patients suffering from the Cotard's syndrome who claim that they are dead or they don't exist (Debruyne, Portzky, Van den Eynde, & Audenaert, 2009), and the patient R.B. reported by Stanley B. Klein and Shaun Nichols (2012), who could recall his memory but claimed that he didn't "own" the memories(p. 685). This sense of identity emerges from a conscious, synchronically coherent ASM. That is, those whose ASMs are not diachronically coherent or globally veridical such as patients suffering from DID or Alzheimer's disease still remain a sense of identity.

There is another sense of identity which refers to the phenomenal experience of being the same as someone in the past or in the future. It results from "self-identification". The concept of self-identification refers to the realization that we are the same individual as an imagined past, future, or possible individual. It requires that the given system possess the capability of forming not only future- or past-models but also one's relation to them. It is worth noting that "self-identification" and "self-connectedness", as introduced by Derek Parfit (1984), are two distinct concepts. (I will address this distinction later in §3.4.4.)

What are the conditions for self-identification? According to SMT (Metzinger, 2004), the concept of phenomenal self-simulation refers to phenomenal representation of a possible state of the system (see §2.2); that is, if another model is successfully integrated into a preexisting and transparent PSM and thus is not

recognized as a model by the system, it is owned by this very system. Accordingly, self-identification arises when a past, future, or possible self-model is integrated into the current self-model, and the content of the given state is "marked as my own".

Several factors are involved in determining degrees of self-identification. There are *trait* and *state differences* in the degree of self-identification. For trait difference, some individuals are more capable of self-simulation than others (D'Argembeau & Van der Linden, 2006). State differences, on the other hand, illustrate the different likelihood of a possible model being integrated into the current state: There are some possible models that are more likely to be integrated than others by the current self-model. The difference is largely dependent on the current state of the system: For instance, compatibility between the autobiographical and emotional contents of the current state and a given possible state. As the systems constantly aim to maximize its coherence, the success of the integration relies on the coherence between the model that is being integrated and the current phenomenal model: A consistent self-simulation is more likely to be integrated, and one is more likely to identify its content as its past self. Another factor is the amount of information available in order to vividly simulate the possible scenario (e.g., the external cues): When one is provided with more information (especially from the first-person perspective), it is easier for the system to construct the relevant context for integrating self-simulation and for identifying with a past self.

The state difference has been illustrated by the property of ASM—diachronic coherence. When one's ASM is diachronically more consistent, one is more likely to integrate past self-model into the current phenomenal ASM. The contrast can be found between a steady person—who lives in a very stable environment and never undergoes dramatic change of personality—and a dementia patient—who generates unreal and often contradictory autobiographical self-model at different times.

3.4.3 The Biological Approach and Transtemporal Human Identity

Traditional discussions of "personal identity" consider the criteria that make a person numerically identical to a person across time. The views involved can roughly be categorized as the psychological and the biological approach. Yet discussions of the biological approach that speaks with the term "person", has neglected considerations of what we essentially are (as noted in Olson, 1997) and to which the concept of person that such approach refers. Either they use the term *person* synonymously with *individual*, *being*, or *us*—instead of the concept of

person we discussed in §3.3.2 and §3.3.4, which refers to a being equipped with psychological traits that allow it to join the moral community—or they claim that "person" is identical to some biological traits. To avoid confusion, in this section (§3.4), I will us use "Person" (upper case P, abbreviation "P") to denote the use of the term by the biological approach which refers to "individual" or the biological concept of person; and "person" (lower case p, abbreviation "p") is reserved to denote the normative concept of person we characterized in §3.3.[21] This section will first consider different versions of the biological approach, which aim to reduce the criteria for transtemporal identity to biological criteria, and then based on animalism, the only meaningful transtemporal relation is human identity.

For those who argue for the biological criteria of Personal identity, such biological criteria stem from two intuitions: First, no psychological characteristics, such as memory, personality, or belief, are unchangeable. Dementia patients at their mid- or late-stage are an extreme example: They can change many of their psychological characteristics dramatically within a short period of time, nevertheless, we are reluctant to consider them as changing from one person to another. Second, if being a Person requires that the subject has some psychological capacity, or that it preserves some psychological characteristics, we may find that one ceases to be a Person when one loses these capacities or characteristics, for instance, patients in a persistent vegetative state (PVS) or in dreamless sleep.

The traditional discussion of Personal identity (transtemporal identity, TI) considers the following question ("Person" in the sense of individual):

> (TI): Assuming that a Person *P* is considered at time *t* and a person *P'* is considered at time *t'*, in virtue of what is Person *P* identical to person *P"*?

According to the biological criteria (BC), the criterion of personal identity is:

> (BC): If *P* is a person at *t*, and *P'* exists as a person at *t'*, then *P* is the same Person as *P'* if and only if *P* is biologically continuous with *P'*.

The concept of biological continuity refers to body, brain, or organism continuity.

Those who endorse the bodily criteria claim that personal identity consists in the sameness of body. They do not preclude the change of cells, since we have new cells growing and old cells dying every day. It is quite obvious that this criterion is

[21] In the entire dissertation, if I use the term "person" without any specification, I refer to the normative concept of person I introduced in §3.3.4.

not sufficient. The corpse after death is an example. Here the body remains the same, yet it is unlikely to be considered a person. Therefore, organism continuity seems to be a more plausible candidate: Personal identity consists in the sameness of a living organism, i.e., the spatiotemporal continuity of a functioning human body. This criterion is what Locke regarded as the sameness of man or plant. It allows the replacement of body parts as long as the functioning of the organism remains the same. If one holds this view, patients in a vegetative state will be considered persons—whereas advocates of the psychological approach claim the opposite: they are not (the same) *person*.

Body transfer is considered an objection to organism continuity. Such ideas exist in Locke's (2008) story of a prince who inhabited a cobbler's body, and also in some films and television series (e.g., *Star Trek*). If it is true that the sameness of living organism constitutes Personal identity, these stories aren't coherent. Another objection is brain transplant. If an individual whose bodily function is extremely weak, has to undergo surgery to transfer his brain into another body or an artificial body, is he still the same Person? Those who defend organism continuity will have to claim that this results in the death of this Person.

Nevertheless, some defenders of the biological approach hold that the person survives the surgery, for they claim that personal identity consists in brain continuity or the continuity of brain functioning. This account is as insufficient as the bodily criteria. But it can still be defended by relaxing the criteria to sameness of a part of the brain. However, some questions have yet to be answered: Which parts of the brain are crucial here? Which brain areas, connections, or functioning are involved?

The most popular biological approach to transtemporal identity is animalism. Following our previous discussion, animalism holds that we are essentially human animals (DeGrazia, 2005b; Olson, 1997; see §3.3.1; Snowdon, 1991), and hence animalists only concern the issue of transtemporal human identity (or animal identity according to Olson, 1997), and are careless with issues of personal identity, since they do not think person is what we essentially are. (Olson denies that there can be a comprehensive criterion for personal identity.)

Transtemporal human identity (THI) concerns the question of what the necessary and sufficient criteria are for a human animal at one point in time to be numerically identical to a human animal at another point in time:

> (THI): Assuming that a human animal *H* is considered at time t and a human animal *H'* is considered at time *t'*, in virtue of what is human animal *H* identical to human animal *H'*?

Then, let's take a look at animalist account of transtemporal human identity.

According to Olson (1997), what determines a being as an animal is its "capacity to coordinate and regulate its metabolic and other vital functions" (p. 133). He discusses a thought experiment in which Tim's body and Tom's head are removed. The question is whether Tim's headless body and Tom's bodiless head remain animals. For Olson (1997) the latter but not the former is still an animal, because the bodiless head remains control of its autonomous nervous system and direct its vital functions, whereas the headless body does not. The bodily head may die in few minutes, but it still remained an animal for a short period of time after being removed from the body, while the headless body is no longer an animal once it has lost connection with the head.

As for the transtemporal identity relation, Olson (1997) contends the following:

> If *H* is an animal at t and *H'* exists at *t'*, *H* = *H'* if and only if the vital functions that *H'* has at *t'* are causally continuous in the appropriate way with those that *H* has at *t*. (p. 135, symbols adjusted by me)

First, an animal is not identical to a body. The body can persist without being an animal. Second, there is no "dead animal". An animal ceases to exist when it dies—when it's vital functions irreversibly stops, and what is left is merely a corpse. However, "causally continuous" and "appropriate way" require further explanation.

In addition, it is required that we return to the beginning of the inquiry. How did the issue of personal identity arise? Why does it attract so much attention from philosophers? What is it that actually matters (Parfit, 1984)? I will discuss this in §3.4.5 and we will see that animalism have failed to respond to any of the issues that matter.

3.4.4 The Psychological Approach and Transtemporal Personal Identity

This section considers the psychological accounts of personal identity. Psychological approaches are distinguished into reductionist and non-reductionist views (Parfit, 1984):

> They are Reductionist because they claim (1) that the fact of a person's identity over time just consists in the holding of certain more particular facts. They may also claim (2) that these facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists. These facts can be described in an impersonal way. (p. 210)

According to reductionist views, theories of persons and personal identity can be exhaustively explained by theories of brains, bodies, or physical or psychological events. Note that psychological events refer to things such as memory and personality in a materialist context, which presupposes that nothing immaterial exists. Note that in the discussions of the general issue of Personal identity, the biological approach also belongs to the reductionist view. On the other hand, the non-reductionism claims that a theory of personal identity cannot be (fully) replaced by a materialist theory. The *simple view* of personal identity, for instance, claims that personal identity is constituted by the sameness of soul, which is a simple, indivisible, and immaterial substance. Advocates of this account adopt substance dualism and believe that persons exist separately and independently from their brains and bodies.

According to the psychological approach, what determine personal identity are psychological capacities or traits. There are different versions of the psychological criterion (PC), but it can be generalized as:

> (PC): If *p* is a person at t, and *p'* exists as a person at *t'*, then *p* is the same person as *p'* if and only if *p* is psychologically continuous with *p'*.

Psychological continuity can refer to memory continuity or quasi-memory continuity.

John Locke (2008) proposes memory as the criterion of personal identity. Locke emphasizes the importance of experience in forming thoughts, opinions, and attitudes. According to his empiricism, the mind is not merely the product of circumstances and conditions. Rather, the mind works up the simple impressions and ideas produced by sense-experience into complex concepts and notions:

> As both an empiricist and a rationalist, Locke regarded people as powerfully shaped by the world around them, but also as free in some degree from both

> animal need and social determination, and thus as capable of determining some of their thoughts and actions on their own. (Seigel, 2005, pp. 155-157)

However, Locke's rejection of "innate ideas" leads to the question of the identity of a person (Seigel, 2005, pp. 88-89). If it is not possible to identify an essential and original core of a being that is unchanged over time, how could a person be the same being as he was yesterday, or even some time ago? How could we say a person is the same in old age as he is in infancy, given all the physical changes that take place across a life? And this leads to further worries about how people could be responsible for their actions or justly subject to reward and punishment if the whole content of their minds was susceptible to changes because of their dependence on experience.

Locke tackles the issue of identity and diverse (which means "non-identity") in *an essay concerning human understanding*. He distinguishes between the identity of substance, between "man", and "person". Consider an atom existing at a particular time and place: It is the same as itself, and as its existence continues, it is still the same. This also applies to compounds, e.g., a mass constituted by two or more atoms. As long as the atoms remain united, the same mass continues to exist. If one of the atoms is removed, or a new atom is added, that mass stops existing.

However, the identity of living creatures does not depend on a mass of the same substance, because an oak tree or an animal remains the same oak tree or animal when a branch of the tree is cut off or some cells of the animal are replaced. In contrast to the mass, which is only a non-integrated cohesion of particles, the oak tree is an organization of parts that enables the whole to receive and distribute nourishment and to maintain its wood and leaves, etc., in which consists its vegetable life.

> That being then one plant which has such an organization of parts in one coherent body partaking of one common life, it continues to be the same plant as long as it partakes of the same life, though that life be communicated to new particles of matter vitally united to the living plant, in a like continued organization conformable to that sort of plants. (2008, Chapter 27 of Book II)

The identities of animals and plants are like the identity of machines: "[I]t is nothing but a fit organization, or construction of parts, to a certain end, which, when a sufficient force is added to it, it is capable to attain" (Locke, 2008, p. 208). This

applies to the identity of the same man if the constantly replaced particles are successively vitally united to the organized body. When we refer to a human as the same being he was yesterday or years ago, it is the organic, corporeal identity we have in mind. Locke understands the term "man" as the idea of an animal of a certain form.

Still, unity of substance does not constitute all forms of identity. The identity of a person is that which exists as a thinking intelligent being, who has reason and reflection, can consider itself as itself, and is the same thinking thing in different times and places (Locke, 2008, p. 208). According to Locke (Locke, 2008, p. 210), consciousness, accompanying thinking, perception, and sensation, makes a subject what he calls a "self", and thereby he differentiate himself from others. The self is:

> […] that conscious thinking thing (whatever substance made up of, whether spiritual or material, simple or compounded, it matters not) which is sensible, or conscious of pleasure and pain, capable of happiness or misery, and so is concerned for itself, as far as that consciousness extends (2008, pp. 210-211).

That is, any part of my body that can be united to consciousness (e.g., the little finger I can feel when touched) is part of my self. Consciousness is an essential component of the self, and is used here in an immediate, reflexive sense that makes the self a "self to itself."

Personal identity therefore depends entirely on being conscious: "For it being the same consciousness that makes a man be himself to himself" (Locke, 2008, p. 214). Thus, as far as the consciousness can be extended backwards to any past action or thought, the self is the same as it was then and personal identity maintains. As Locke's example shows, if the consciousness of a prince occupied a cobbler's body, it would be the same person as the prince instead of the cobbler. It is important to note that Locke used the term "consciousness" to refer to knowledge about one's actions as well as one's more general knowledge.

The term "person", as Locke uses it, is a forensic term for "self" (p. 211). Personal identity is the basis of rights and justice, namely of reward and punishment. If someone who does not partake the same consciousness as he did when he was sleeping, it would be unjust to punish him awake for what he did asleep, since he is not the same person.

Joseph Butler and Thomas Reid have proposed objections to memory criteria as well as to Locke's view on personal identity. In favor of a substance-based view of identity, the main objection from Butler (2008) is that consciousness (and memory) presupposes personal identity:

> And one should really think it self-evident, that consciousness of personal identity presupposes, and therefore cannot constitute personal identity, any more than knowledge, in any other case, can constitute truth, which it presupposes. (p. 100)

That is, I can remember my experience, but it is not the memory that makes it mine, instead, I can remember because the memory is already mine.

Thomas Reid (2008) adds that such criteria allow situations that violate the transitivity of the identity relation. Consider the case of a brave officer who has been flogged for robbing an orchard as a boy, who later stole the standard from his enemies, and became a general in his advanced life. This situation is possible: The officer remembers the flogging when he was stealing the standard; nevertheless, after becoming the general, he can remember the standard stealing but has lost his consciousness of the flogging, or the other way round (remember the flogging and forget the standard stealing):

> These things being supposed, it follows, from Mr. Locke's doctrine, that he who was flogged at school is the same person who took the standard, and that he who took the standard is the same person who was made a general. Whence it follows, if there be any truth in logic, that the general is the same person with him who was flogged at school. But the general's consciousness does not reach so far back as his flogging; therefore, according to Mr. Locke's doctrine, he is not the person who was flogged. Therefore the general is, and at the same time is not, the same person with him who was flogged at school. (pp. 114-115)

These objections lead to Sydney Shoemaker's proposal of quasi-memory (q-memory). Q-memory is a weaker sense of memory that presupposes no personal identity:

> Whereas someone's claim to remember a past event implies that he himself was aware of the event at the time of its occurrence, the claim to quasi-

> remember a past event implies only that someone or other was aware of it. (Shoemaker, 1970, p. 271)

Accordingly, I can quasi-remember or have a q-memory of something that I did not do or didn't experience. The concept of q-memory not only answers Butler's objection about memory's presupposition of personal identity, but also provides a way out in the case of the brave officer. Based on "q-memory criteria", the old general is the same person as the flogged boy, since the old general remembers the standard-stealing officer who in turn remembers the flogging. In other words, the old general q-remembers the flogging and thus is the same person as the boy. However, this modification, in the form of q-memory, cannot explain the fact that there are some points in our life at which we cannot remember or q-remember at all, like in dreamless sleep and vegetative states, for instance. Does it mean that I am not the same person as the one who slept on my bed last night?

Another objection to psychological criteria (criteria for numerical identity in general) is the *fission problem* (Olson, 2003b, p. 361; Parfit, 1984, pp. 254-255). To begin with, if one person p has his cerebrum transplanted, advocates of the psychological criteria may argue that the being with the transplanted cerebrum is the same person as p, since she is psychologically continuous with p. (The premise that the cerebrum carries the psychological continuity may be problematic.) If one of the hemispheres of person p' is destroyed, and the other is transplanted, the person inheriting the hemisphere is the same person as p'. That is, personal identity maintains in the process of transplanting half of the cerebrum, and the cerebrum carries the identity. Now p undergoes a surgery that transplants each of his hemispheres into two different bodies, and this procedure results in two beings Lefty and Righty. (We assume that there is no dramatic difference between the two hemispheres.) There are three possibilities concerning personal identity in such a situation: (1) The person stops surviving; (2) the person survives the fission as one of Lefty and Righty; (3) the person survives as both of them.

At first glance, the second option will have some difficulties: Since we assume that there is no decisive difference between the two transplanted hemispheres, it will be difficult to tell which of them inherits the identity. Also, the third option is not plausible, for if both of them continue to be the same person as p, it follows to be that p has numerical identical with Lefty and Righty. But numerical identity is a reflexive, symmetrical, transitive and one-one relation. Psychological

continuity, which may result in a one-many relation between a person and others in the past or future, cannot be the criteria of personal identity.

The double occupancy view is provided to avoid this problem (as cited in Olson, 2003b, pp. 361-362). The former argues that there are actually two people occupying the same place, who are made of the same matter before the fission. What the surgery does is to separate two people. This view is usually suggested in the context of four-dimensionalism.

The non-branching view (NBV), taking the first option, considers the criterion of psychological continuity insufficient for personal identity. The psychological criterion is therefore modified:

> (PC_{NBV}): If *p* is a person at *t*, and *p'* exists as a person at *t'*, then *p* is the same person as *p'* if and only if *p'* is *uniquely* psychologically continuous with *p*.

Accordingly, the person cannot survive the fission because for a person, if there exist more than one person that is psychological continuous with her, she ceases to exist. Thus, a person will not be able to survive the duplication either.

I have reviewed different versions of reductionist personal identity, but no one approach is satisfactory. What is the relation of personal identity? The possibilities are as follows:

1. There is such a relation in reality that corresponds to personal identity,
   … and the criteria can be found through reductionism.
   … and the criteria cannot be found through reductionism.
2. There is no such a relation in reality that corresponds to personal identity.

Most philosophers have regarded the relation of personal identity as something that exists in reality (1), such as the relation between parental generation and offspring, which can be determined by a factual criterion such as genotype. The criteria of this kind of relation can be found in reality. Does personal identity belong to this kind of relation? Can we find a factual criterion of personal identity?

In our review of reductionist accounts, none of the accounts from psychological or biological approaches seemed plausible, for they all encountered their own problems and objections, which were left unsolved. Besides, one of the main objections stems from the fission problem, mentioned as an objection to the

psychological approach. Yet this haunts not only psychological approaches but also biological ones: The advocates of these reductionist approaches have tried to come up with a property or a set of properties as the criteria of personal identity, but these "property-based approaches" are always problematic, since a scenario in which there are more than one person who has those properties at another time can be easily conceived, thus the relation of "identity" is violated. Unless we added uniqueness to the criteria, this situation cannot be avoided. Nevertheless, we have discussed why the criterion of uniqueness is undesirable as well. (In fact, the identity relation of anything may face "the fission problem" as long as its criteria are "property-based"—not merely the identity relation of persons.) Thus, it seems that the reductionists (1a) have failed to offer a solution to the problem of personal identity.

We are left with two options: We have either to accept that the criterion of personal identity cannot be found through a reductionist approach (1b), or we acknowledge that personal identity has no ontological status (2). If the former is correct, we may end up accepting dualistic metaphysics or something non-physical. I argue that we cannot find a descriptive criterion for transtemporal personal identity (2): The problem is a normative problem. The relation of personal identity is not like the genetic parent-offspring relationship, where there is a metaphysical truth; instead, it is a matter of conventions, norms, and personal and social opinions, based on the normative concept of personhood.

The idea of transtemporal personal identity is built upon the idea of personhood: If one is not qualified as a person, the identity relation does no longer hold. Therefore, the normative concept of personhood is significant in the consideration of the transtemporal personal identity. The question of what a person is has been discussed in §3.3.4 and we concluded that "person" is a normative and dynamic concept. That is, the concept of "person" is close to the concept of a "citizen" in two senses: (1) The criteria of being a person or a citizen cannot be found in nature. They are not descriptive but normative. (2) They may change overtime and vary from culture to culture according to our conventions, norms and preferences.

Accordingly, if we want to understand the relation of transtemporal personal identity, it is close to the way in which we judge the identity of a citizen. There are two parts involved: a normative part that determines what the criterion is (the decision as to what the criterion of being the same citizen is) and a factual part (the determination of the same identification card number or the same fingerprint). The

normative aspect of the criterion may differ from a moral community to the other. Like the concept of personhood, it relies on the shared way in which the members of the moral community conceive each other as a moral agent that exists diachronically. This results in how we interact with each other. Then, in order to determine what are involved in the criteria within one moral community, the empirical investigation (e.g., on the degree of satisfaction of the diachronic coherence of the ASM) is required.

3.4.5 The Moral Significance of Transtemporal Identity

Because it seems impossible for transtemporal identity to avoid the problem of fission mentioned above, Parfit (1984) proposes an alternative: A view that disregards identity and holds that what really matters in one's survival is relation R. Parfit (1984), while sympathetic to psychological criterion, recognizes that it seems like this criterion of personal identity cannot avoid branching, which is impossible for the relation of identity. Adding uniqueness into the criteria may solve the problem, as in this case the person stops surviving. Nevertheless, this idea of not surviving is far removed from our ordinary idea of death. If a person is duplicated, it makes no sense to declare the end of a person's existence just because there is another person who is psychologically similar to him. The person did not die: He is still alive with a personality and memory. Thus, Parfit (1984) suggests the following:

> We might say: 'You will lose your identity. But there are different ways of doing this. Dying is one, dividing is another. To regard these as the same is to confuse two with zero. Double survival is not the same as ordinary survival. But this does not make it death. It is even less like death.'
>
> The problem with double survival is that it does not fit the logic of identity. Like several Reductionists, I claim
>
> > Relation R is what matters. R is psychological connectedness and/or psychological continuity, with the right kind of cause. (p. 262)

As we have seen, there are two components of the psychological criterion of personal identity (PI): psychological and/or psychological continuity (R) and uniqueness (U). However, the presence or absence of U does not alter the value of R. As for R, psychological connectedness and continuity are distinguished: The former refers to holding of direct psychological connections, whereas the latter is the holding of overlapping chains of direct connectedness (1984, p. 206).

It is worth mentioning that psychological connectedness is a distinct concept from the self-identification introduced in §3.4.2: First, they are involved in two different questions suggested by Marya Schechtman (1996). Psychological connectedness responds to the reidentification question, which concerns whether an individual at a time is the same person as an individual at another time. On the other hand, self-identification is involved in the characterization question, which considers the conditions under which various psychological characteristics, experiences, and actions are properly attributable to a person. Second, the degree of self-identification and psychological connectedness can differ. This is especially evident in cases of confabulation and self-deception. For instance, we have seen that RZ, a 40-year-old woman, had the delusional belief that she was her father or occasionally her grandfather (Breen et al., 2000). When she had the delusion, RZ had a strong self-identification to her imagined past self, but according to Parfit (1984) only a weak psychological connectedness with her "ancestral self". GA, a 52-year-old woman, as another example, developed the amnesic-confabulatory syndrome and often made implausible future plans (Barba, Cappelletti, Signorini, & Denes, 1997). For example, to the question "What are you going to do in a few minutes?", GA once answered "I will go home to cook the supper" (p. 430). Since developing her disease, GA had actually never cooked and was living a hospital an hour and a half away from her home. For GA, there is strong self-identification with an implausible self-model, but weak psychological connectedness to an upcoming or "descendent self".

Returning to the issue of what matters in personal identity and recalling the normative concept of person we discussed in §3.3.4, what is important in the discussion of personal identity is what a person is. Personhood, as I suggested earlier, is a property of the self-model that has fulfilled the criteria required to become a member in the moral community. The requirement of being a person alters according to the mutual conception of the members. Consequently, the criterion of transtemporal personal identity can also change: It will change with the concept of person, and together their changes lead to different ideas of rights and responsibilities.

I agree with Parfit that what we are looking for in the discussion of transtemporal personal identity—the relation that provides moral implications to important practical issues—is psychological connectedness and continuity. Following Parfit's proposal of relation R, what is decisive in transtemporal personal

identity (or transtemporal personal relation[22]) is the relationship between ASMs—that is, collections of mental self-simulations about the contents of one's past or potential future PSMs constructed at different times. The degree of psychological connectedness is dependent on the structural similarity between the ASMs. This can be illustrated by the diachronic coherence of an ASM: A stronger psychological connectedness exists when the ASMs constructed are more diachronically coherent. The property of diachronic coherence may provide some moral implications and insights to the consideration of normative issues. For instance, whether an advance euthanasia directive made before advanced dementia can decide for the patient (see §8.4.2), relies on the degree of diachronic coherence.

3.5 Summary

This chapter reviews the concepts of selfhood, personhood and personal identity. First, based on the SMT (Metzinger, 2004), phenomenal self and subjectivity are explained by the satisfaction of functional constraints (§3.1).

- *Different concepts of the self are endorsed: The narrower concept refers to the phenomena which are accounted by the self-model and the functional constraints, whereas the richer concept includes the contents of the PSM and ASM.*
- *There is neither any empirical evidence to support the ontological existence of a self, nor any conceptual argument for the necessity of ontological realism of the self.*
- *How a self-model becomes conscious is explained by the satisfaction of constraints of globality, presentationality, and transparency (§3.1.4). This kind of PSM doesn't recognize itself as a subject until further constraints—convolved holism, dynamicity, and perspectivalness—are satisfied (§3.1.5). A PSM thus allows the emergence of subjectivity.*

Next, the relations between different kinds of memory (i.e., working, episodic, and semantic memory) and self-model are discussed (§3.2.1). After reviewing the autobiographical self (Damasio, 1999; 2010; §3.2.2) and the ego dysequilibrium theory (Feinberg, 2005; 2009a; 2009b; §3.2.3), the ASM is introduced.

[22] To avoid the problem of branching discussed in the last section.

- *Working memory is necessary for the emergence of a PSM: As a limited capacity for storage, it makes a PSM possible by providing a workspace, which holds the phenomenal representations temporarily active. PSM, on the other hand, by integrating the (self-)simulations constructed by episodic and semantic memory allows the memory (self-)simulations to be cognitively and consciously accessed.*
- *An ASM is a collection of mental self-simulations of the relations with past and potential future states. Its contents can be cognitively and consciously accessed when it is integrated into the currently active PSM.*
- *Three constraints are involved in the construction of an ASM: Synchronic coherence refers to the consistency of the contents of (self-)simulata of the ASM constructed at a particular time; diachronic coherence refers to consistency between the contents of ASMs constructed at different times; and global veridicality refers to how faithfully an ASM simulates past experience and events.*

Then, in order to prepare for later discussion on the issue of authenticity, the question of our fundamental nature and the issue of transtemporal identity are discussed. I suggest that there are different kinds of transtemporal relation—transtemporal personal relation and transtemporal human relation. Depending on the different concepts of human animal and persons (§3.3.1 & §3.3.4), they have different criteria for identity (§3.4.3 & §3.4.4)

- *Transtemporal human identity refers to how a human animal at one time is identical to a human animal at another time; transtemporal personal identity concerns how one person at one time is identical to a person at another time. The biological approach accounts for transtemporal human identity*
- *The transtemporal personal identity relation is based on personhood, which I argue is a normative concept: It is the property of a whole system that emerges when the requirements of membership of the moral community are fulfilled. Accordingly, personal identity is normative: It relies on how members of the community conceive each other as moral agents that exist diachronically.*

Chapter 4

Conceptual Tools III: Enhancement

4.0 Introduction
4.1 Utilitarianism and Suffering
4.1.1 Negative Utilitarianism
4.1.2 Suffering
4.2 The Conceptual Issues of Enhancement
4.2.1 Definitions of Enhancement
4.2.2 The Function of the Distinction and the Health-Based Concept of Enhancement
4.3 The Concept of Health
4.3.1 Normality and Normalization
4.3.2 Christopher Boorse on Biostatistical Theory
4.3.3 Lennart Nordenfelt on the Holistic Theory of Health
4.4 The Phenomenological Account of Health and Enhancement
4.4.1 The Reverse Theory of Disease and Illness
4.4.2 The Phenomenological Account of Health
4.4.3 The Phenomenological Account of Enhancement
4.5 Summary

4.0 Introduction

In order to bring forth a clear definition of "memory enhancement", I shall need to define "memory" and "enhancement". I have delineated the concept of memory in Chapter 2; here I propose a definition of enhancement, which I suggest is more adequate than currently available notions. Later, in Chapter 6, based on the conceptual analysis on "memory" and "enhancement", I will suggest an account of memory enhancement.

The increasing power to manipulate human cognition by intervening in brains and the growing application of pharmaceuticals, designed to treat mental illness, on the healthy for enhancing cognitive functions have generated a great deal of ethical debates. Although the issue is often raised, and the concept of

enhancement is confronted in both academic publications as well as in the mass media, the concept remains vague:

> It is often assumed that the distinction between therapy and enhancement is clearly defined; similarly with the distinctions between normality and abnormality, and health and disease. The assumption also seems to be made that whatever these distinctions may be, they are unchanging. They are the same today as they were in 1950 or 1850, or they are the same now across all societies. (Jones, 2006, p. 78)

There have been different understandings of "enhancement". For instance, Bostrom and Sandberg (2009) have defined cognitive enhancement (CE) as "the amplification or extension of the normal cognition through improvement or augmentation of internal or external information processing systems" (p. 311). Metzinger and Hildt (2011), on the other hand, define it as a technology that "aims at optimizing a specific class of information-processing functions: *cognitive* functions, physically realized by the human brain" (p. 245).

As Jones (2006) and Metzinger and Hildt (2011) suggest, two distinctions are crucial for the concept of enhancement:

- **The distinction between health and disease/illness:** "Enhancement" often refers to application on healthy individuals; therefore, without clearly identifying what health and ill-health mean, the concept of enhancement remains ambiguous.
- **The concepts of normality and abnormality**: Both the distinctions between treatment and enhancement, and between health and ill-health involve the idea of normality, e.g., "normal functioning", "normal/standard circumstance". What is normal and what is abnormal? It will be helpful to understand how the concept of normality is used in a variety of ways.

I will argue for a treatment-enhancement distinction that is compatible with medical concepts such as health, disease, and illness. As for the concept of normal, as we will see in §4.3.1, there are at least seven conceptions of "normal". I contend that it could help us understand how the relevant notions are conceived, but there is no use and no need to distinguish people or conditions with any notions of normal.

In this chapter, I will introduce negative utilitarianism and suffering, which will be the theoretical foundation of my demarcation of "health" and "illness" as well as "enhancement" and "treatment" (§4.1). Next, I review current approaches of defining "enhancement" and propose the health-based concept of enhancement (§4.2). Then, by means of different ways of understanding normality differentiated by Edmond A. Murphy (1966), more precise conceptual tools are acquired (§4.3.1). To introduce the debate between objectivism and constructivism about health, I focus on the value-free Biostatistical theory of health developed by Christopher Boorse (1975, 1977) and value-laden Welfare theory of health given by Lennart Nordenfelt (1993, 2001, 2007) (§4.3.2–§4.3.3). Finally, I propose a phenomenological concept of health (§4.4.2), and the account is used further to define the concept of enhancement (§4.4.3).

4.1 Utilitarianism and Suffering

This dissertation takes a moderate version of negative utilitarianism as the default practical theory. It is *practical*, because my proposals in the dissertation are not restricted to any particular ethical theory. Moderate negative utilitarianism is used as a theoretical background for two minimal criteria: the minimization of suffering and self-determination. This section aims to set the theoretical ethical ground. Before defending negative utilitarianism, suffering is introduced.

4.1.1 Negative Utilitarianism

There are different forms of utilitarianism. *Act utilitarianism* holds that the right action is the one that has the best consequence of all actions open to the agent; on the other hand, according to *rule utilitarianism*, the right action is the action that follows the rule that results in the best consequence of the available rules. But how is the consequence considered? What exactly is taken into consideration? *Classical or hedonistic utilitarianism* considers pleasure and pain: The only intrinsically good and bad things are, respectively, pleasure and suffering.

However, objections to hedonistic utilitarianism include that suffering cannot be compared to pleasure. Suffering has a distinct moral weight that pleasure does not have. The moral significance of suffering is neatly expressed in the following passage from Peter Singer's (2011) *Practical Ethics*:

> If a being suffers, there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the

> principle of equality requires that the suffering be counted equally with the like suffering—in so far as rough comparisons can be made—of any other being. If a being is not capable of suffering, or of experiencing enjoyment or happiness, there is nothing to be taken into account. This is why the limit of sentience is the only defensible boundary of concern for the interests of others. To mark this boundary by some characteristic like intelligence or rationality would be to mark it in an arbitrary way. Why not choose some other characteristic, like skin color? (p. 50)

That is, being capable, i.e., being prone to suffering, not pleasure, makes sentience the basis of moral concern and perhaps the only non-arbitrary boundary of concern.

This view leads to *negative utilitarianism*. In contrast to happiness-maximization—the principle of hedonistic utilitarianism—the principle of negative utilitarianism is suffering-minimization. Karl Popper (1971) proposes this version of utilitarianism in his classic work, *The Open Society and Its Enemies*, contending that one of the most important principles of humanitarian and equalitarian ethics is

> [t]he recognition that all moral urgency has its basis in the urgency of suffering or pain. I suggest, for this reason, to replace the utilitarian formula 'Aim at the greatest amount of happiness for the greatest number', or briefly, 'Maximize happiness' by the formula 'The least amount of avoidable suffering for all', or briefly, 'Minimize suffering'. Such a simple formula can, I believe, be made one of the fundamental principles (admittedly not the only one) of public policy. (The principle 'Maximize happiness', in contrast, seems to be apt to produce a benevolent dictatorship.) We should realize that from the moral point of view suffering and happiness must not be treated as symmetrical; that is to say, the promotion of happiness is in any case much less urgent than the rendering of help to those who suffer, and the attempt to prevent suffering. (Popper, 1971, vol. 1, chapter 5, note 6)

Therefore, according to negative utilitarianism, the right action is the one that will, in the long run, minimize the total amount of suffering (Kadlec, 2008).

However, negative utilitarianism requires further constraints if one wishes to avoid the following problems: First, a radical version of negative utilitarianism without further constraints will lead to the problematic implication that if some sort of suffering is not preventable in one's life, it is better we have never existed (Benatar, 2006). Then, we ought to avoid having children, painlessly end everyone's

life, and let human beings extinct because these approaches efficiently meet the criterion of minimizing overall suffering. It could also imply that bringing more new beings into the world is a wrongful act since it produces more suffering? Second, there are always those who are willing to endure suffering in order to reach certain goals. For instance, Pierre-Auguste Renoir continued to paint despite the fact that he suffered from rheumatoid arthritis. Labor pain is another example: Labor is notoriously painful and involves suffering, but many women are willing to go through it for the sake of having children.

As we will see in the next section, suffering develops to urge the system to escape from the current state. However, as human beings or other animals have developed cognitive functions (e.g., memory and reasoning) which allow them to be detached from the current situation spatially and temporally and to have culture and appreciation of abstract objects (e.g., art, friendship, and justice), the limitations suffering imposes on the organism could be lessened, and one can pursue goals in a more distant future or one that does not involve current suffering. The upshot is that suffering has the intrinsic feature that we strive to minimize it, but this determined by one's interests when the organism is equipped with the capacity to distance itself from the occupation of suffering.

These cases challenge negative utilitarianism, because despite the great pain a person might be suffering, we contend that their own interests should be respected. The idea of preference satisfaction is the central thesis of *preference utilitarianism*, according to which the right action is the one that will, in the long run, satisfy more preferences of those who are affected than it will thwart (Singer, 2011, pp. 13-15). I sympathize with preference utilitarianism; however, I have not chosen to adopt it here for two reasons: First, not all kinds of preferences should be considered equal. Some preferences, such as those that relate to suffering and survival, should be treated as more urgent than others. However, if there indeed are more important preferences, what makes some preferences more morally significant than others? Without any additional constraints, the criterion of the moral significance of a preference again relies on preference of preferences. There seems to be an infinite regress. Second, here I only want to provide a minimal set of moral constraints without committing to a specific moral theory. In my opinion, negative utilitarianism is less demanding than preference utilitarianism and can be compatible with other moral theories.

To return to negative utilitarianism, what are the further constraints we need to add in order to allow for suffering that in turn allows something else to be

respected? Metzinger (2013), adopting a moderate version of negative utilitarianism, also emphasizes the importance of self-determination:

> Moderate versions of utilitarianism respect individual rights, and it seems to be a very widespread and fundamental intuition within human societies that the preference for existence and self-determination must not be frustrated without good reason for any self-conscious entity which possesses it. (p. 270)

I contend that the constraint of suffering-minimization should be based on one's right to self-determination. Any action to minimize suffering should not contradict the preference of those whose suffering is reduced. That is, according to this adjusted version of negative utilitarianism, an action is right if (1) the action minimizes overall suffering, and (2) those whose suffering is reduced prefers the result of the action. Consequently, if a patient suffers from great pain resulting from an incurable disease, it is the right action to end her life if she prefers it; nevertheless, it is not a right action if she does not have such a preference.

This version of moderate negative utilitarianism provides two constraints: That the right action should reduce overall suffering, and the action should be based on the consent of the one whose suffering is minimized. These two constraints seem to be independent, but in fact, they rely on the nonexistence of suffering. In the next section, I will briefly introduce suffering and its relation to preference.

4.1.2 Suffering

Suffering is a phenomenal event which is available only to sentient beings—beings with the capacity to experience. It is distinct from pain and stress, whereas it can be the result of them as well as anxiety, fear, boredom, and a variety of stimuli (DeGrazia & Rowan, 1991, p. 199). While most philosophers have agreed that animals do experience pain[23], whether animals or infants suffer is rather debatable. Jamie Mayerfeld (1999) defines suffering as a disagreeable *overall* feeling. Here, "overall" is emphasized is to underscore the characteristic of suffering that it is a negative condition of the system as a whole. One can go through pain and joy—what is presumably bad and good—simultaneously; however, it is not the case for the condition like suffering. Local tissue damage, for instance, directly leads to pain, while it does not necessary result in suffering. What makes one suffer does not

[23] Descartes, who regards animals as automata, is an exception.

necessarily make another suffer. The occurrence of suffering is associated with biological, psychological and social conditions. How the current situation is evaluated by the subject is central to whether stress or pain leads to suffering. For instance, an athlete finishing a marathon experiences pain and all sorts of unpleasant feeling, but the evaluation of his performance may prevent her from suffering.

Eric Cassell (2004) understands suffering as a "state of severe distress associated with events that threaten the integrity of person" (p. 32). Cassell contends that only persons can suffer because according to his view, the occurrence of suffering as well as the to which degree it happens, depend and vary in accordance with attitudes or expectations of subjects. The belief that suffering will soon be removed certainly can lessen the intensity. That is, Cassell indicates the importance of "meanings" as critical determinants of whether and how one suffers: how one assigns meaning to one's situation in relative to one's mental autobiography (the expectation of the future in particular). Due the central role of meaning in suffering, beings without meanings, "temporal self-awareness" or advanced psychological capacities, such as having concepts, are incapable of suffering. Only persons—defined in terms of complex psychology capacities—are capable to suffering.

This view of suffering is concurred by Daniel Dennett and Michael Tye. Dennett (1996) ascribes the lack of suffering to the absence of the sort of informational organization in animals. "The anticipation and aftermath, and the recognition of the implications for one's life plans and prospects" (p. 167) constitute suffering. As a result, in order to understand suffering, as Dennett maintains, study of the subject's life instead of its brain is required. Tye (2000) argues for a similar view that suffering requires cognitive awareness of one's pain, and as a result, animals unable to form cognitive awareness is free from suffering.

Other philosophers object these views and argue for the possibilities of animal suffering (e.g., DeGrazia, 2014). Metzinger (2013, pp. 263-266) proposes a "minimal model of suffering", with four necessary conditions for a being to suffer—C-, PSM-, NV-, and T-conditions. The C-condition states that suffering is a *phenomenological* concept, and only beings with conscious experience can suffer. According to PSM-condition, a conscious being without a coherent PSM cannot suffer. If the system does not integrate negative states into its PSM, one cannot have the sense of ownership of suffering. Then the NV-condition emphasizes the negative value (,which the system does not prefer and will struggle to escape from) integrated into the system: It urges the system to leave the current conscious state.

Last, the T-condition is that suffering is possible if only it is a phenomenally transparent state. The T-condition makes suffering "irrevocably real", where one cannot distance herself from it. These are the necessary conditions of suffering: any beings—including animals and robots—which satisfy these conditions are capable of suffering.

Marvin Minsky (2006) investigates the relation between pain and suffering. The later further illustrates suffering as a cascade of mental change that disrupts one's plans and goals by suppressing other resources. By doing this, the system is forced to focus on the only interest left: "*Get rid of that pain*" (Minsky, 2006, p. 65):

> The primary function of Pain is to compel one to remove what is causing it—but in doing so, it tends to disrupt most of a person's other goals. Then, if this results in a large-scale cascade, we use words like Anguish or Suffering to describe what remains of its victim's mind. (p. 70)

That is, suffering may have the function of warning and compelling the subject at the conscious level to take action to free itself from an extremely unpleasant situation. Suffering may become more intense as the situation worsens or the same negative state has gone on for too long. It is predicted that neuroscientists may be able to locate the neural correlates of suffering and understand the underlying mechanisms, including the internal neural processing and the external influence of the (social) environment.

I side with Metzinger and Minsky on the issue that temporal or reflective form of self-awareness is not prerequisite for suffering. I contend that suffering emerges when the being is in a kind of mental state (or a form of cognitive architecture). The central characteristic of this sort of mental state is that it urges the system to withdraw from the current state. It occupies all the cognitive resources available to make it happen. I hypothesize that suffering undermines working memory. Because of the reduction of the working space a phenomenal self-model requires, the organism is temporarily deprived of the capacities to keep sufficient information active and to manipulate them for a future goal. Long-terms goals are consequently suppressed and the most urgent and only mission is to modify its system physically—e.g., by relieving the pain that results in suffering—or psychologically—e.g., distracting oneself from a traumatizing recall, that is, modifying one's ASM. The system is confined to escaping the current situation.

Accordingly, when undergoing suffering, the autobiographical self-model (ASM) that is incorporated into the current phenomenal self-model (PSM) is distorted, and is restricted to information that is very closely related to the current situation. Besides, it is predictable that the ASM that is incorporated into PSM under suffering is distinct from that which is incorporated when not suffering: As I discussed in the last chapter, the constructed phenomenally-experienced mental autobiography is not only determined by the mental self-simulations constituting ASM, but also determined by the current goal and situation of the subject. As the goal under suffering is extremely narrow, one is likely to experience very distinct mental autobiography when the suffering is lessened: One will feel like a different person to the one who was suffering.[24]

As such, if one is suffering, one is less capable of determining one's long-term interests. Because the interests one expresses when suffering are likely to be distinct from those one has when freed from suffering, there exists only weak psychological connectedness (Parfit, 1984, p. 206). If there exists only very weak psychological connectedness, but one's decisions for the future are still effective, the problem of self-paternalism arises. A more extreme and perhaps more typical case is advanced directives for euthanasia at a certain stage of dementia. I contend that without strong psychological connectedness through memory and other psychological characteristics, one is incapable of making decisions about the future.

This leads to the practical issue: Can a patient who is experiencing suffering make long-term decisions for herself, such as undergoing voluntary euthanasia? This depends on (1) whether the suffering has undermined the capacity to make judgments and (2) what kind of decision one has to make. First, as discussed earlier, suffering results from the occupation of one's mental resource and working memory, which forces the system to deal with the current problem. This can lead to decreased free working space for the consideration of one's self-interests and long-term plans. If one is undergoing a great suffering which already affect one's capacity of making judgment, decisions should be made later when one's mental resource is freed from such occupation. Second, if the decision is irreversible once made, such as euthanasia, one should avoid making such kind of decision when suffering.

[24] It is interesting that suffering affects one's construction of ASM, but at the same time, Cassell and Dennett are right that whether one suffers strongly depends on one's ASM—how the suffering is related to one's beliefs and value. Even though ASM could affect the occurrence of suffering, it doesn't imply that ASM is necessary for suffering.

However, such consideration is based on the assumption that the patient can be released from suffering in her future. Whether patients under suffering can make decisions for themselves also relies on whether they can be liberated from the suffering. If that which the patient is suffering from can be removed, and in the near future, she can be freed from suffering, and thus her options should be preserved and any decision should be made at a later time when she has the sufficient mental capacity to judge. Even if the disease or disorder is not curable, the decision should be made under the circumstance when the agent is best at making judgments.

What if one cannot escape from suffering even with medical aids? First, having "interests" and "preferences" does not necessarily require complicated cognitive capacities. They are required for organisms such as human beings (or other animals), because of our capacity to make long-term plans and complicated thoughts, which result in greater behavior flexibility, one's interests are less predictable and are more likely to change over time. In contrast, organisms with simpler minds are still capable of having interests based on the limited cognitive capacity required to live. Their interests in living are as real as those of others. Then, those who experience long-term suffering and those who inherently lack or have less capacity for judgment also have interests. Their interests are to be respected as much as our interests. Therefore, any action to reduce their suffering (e.g., to end their life) should respect their interest in survival.

In this chapter, the moderate version of negative utilitarianism is examined in order to show the moral significance of both suffering and self-determination. In the following sections, suffering will be treated as the demarcation for distinguishing health from illness, which determines the distinction between treatment and enhancement. Self-determination will play a role in defining the concepts of (memory) enhancement and enhancer in the next chapter.

4.2 The Conceptual Issues of Enhancement

Although the concept of enhancement and the distinction between enhancement and treatment have intuitive appeal at first glance, on closer inspection both are far from clear. In general, "enhancement" is used in contrast with "treatment": It refers to interventions designed to produce "improvement" in humans, but is not categorized as treatment. Different definitions of enhancement have been proposed (Juengst, 1998, pp. 32-37; Savulescu, Sandberg, & Kahane, 2011). This section introduces

these definitions and addresses the function and significance of the distinction. I will argue for a *health-based concept of enhancement*.

4.2.1 Definitions of Enhancement

The first definition of enhancement, from "disease-based accounts" (Juengst, 1998; or the "not-medicine" approach in Savulescu, Sandberg, et al., 2011), is commonly adopted in discussions of the ethical implications of enhancement (e.g., De Jongh, Bolt, Schermer, & Olivier, 2008; Glannon, 2011; Hildt, 2013). According to Eric T. Juengst (1998), enhancement refers to "interventions designed to improve human form or functioning beyond what is necessary to sustain or restore good health" (p. 29). Treatment and enhancement are distinguished by the problems to which they respond: Treatment is intervention that addresses health problems resulting from diseases and disabilities; enhancements are interventions aimed at healthy systems and normal traits. For instance, medical intervention for curing hepatitis B is, without question, considered a treatment; a rhinoplasty designed to make a nose thinner and more pointed is commonly regarded as enhancement. This mode of distinction is intuitively appealing and consistent with biomedical classification: If there is no pathology identified, the intervention goes beyond treatment. Accordingly, the distinction relies heavily on the concept of disease or abnormal condition. It leads to a question about what disease and health actually are, which we shall look into in §4.3.

Some maladies are surely more straightforward to diagnose, such as influenza, cancer, or acquired immunodeficiency syndrome (AIDS); nevertheless, some are not as clear. For instance, medical intervention for obesity is considered treatment, while intervention on a subject with "normal" weight is enhancement. But here, what is the criterion for obesity? The World Health Organization (WHO) has assented that a person is regarded as obese when her body mass index (BMI), her body mass divided by the square of her height, is over 30 kg/m^2 (WHO consultation, 2000). However, this number is not rigid. A WHO expert consultation (2004) has suggested that Asian countries should have a lower cut-off point for what is considered overweight, because Asian populations develop negative health consequences (e.g., type 2 diabetes and cardiovascular disease) at a lower BMI than Caucasians. In addition, cognitive and mental disorders are especially difficult to identify. Depressive disorder, for instance, has been included in *Diagnostic and Statistical Manual of Mental Disorders* (DSM; American Psychiatric Association, 2013). However, the definition remains controversial: The latest (fifth) edition of

DSM has modified the definition by eliminating "bereavement exception" from the guidelines for diagnosing major depressive disorder (as cited in Bryant, 2014). Prior to this edition, bereavement exception had been in the guideline because grieving after bereavement is regarded as a *normal* response. This issue then triggers further questions, such as what kind of response should be considered *normal*. This dispute over the definition of depression is one of many controversial cases of defining some mental disorders (e.g., autism spectrum disorder and attention deficit/hyperactivity disorder).

Despite the conceptual and practical obstacles to identifying disease, the problem of this account of distinguishing enhancement from treatment includes the demarcation of prevention. Prevention is often automatically considered treatment and accepted as a legitimate part of biomedicine, for example, such as Bacillus Calmette–Guérin (BCG), a vaccine against tuberculosis. Prevention as a treatment is controversial, for prevention implies the non-existence of the disease. One can modify the definition of treatment so that it refers to any intervention targeting an occurrent or dispositional disease. Nevertheless, this expansion of the definition leads to other questions: What is the domain of dispositional disease? Does it include all the disease we can possibly have? Can one always specify the disease or pathology that prevention targets? For instance, strengthening the body's immune system through healthy diets and adequate sleep, or avoiding infection through hygienic habits is general preventions from disease. Should they also be regarded as treatment?

Second, H. Tristram Engelhardt (1984; 1990, cited in Juengst, 1998) regards the distinction between enhancement and treatment as a social construction reflecting the values and willingness of medical professions to perform interventions. The boundary of a practitioner's obligation is based on the value-systems of the patient and practitioner, which include their ideas of "disease", "health", "treatment", "enhancement", "human flourishing", and so on. Through their negotiation, the practitioner and patient together decide which problems count as "disease" and which interventions count as "treatment". Consequently, these concepts are different from person to person and from case to case. The advantage of this account is that practitioners can easily make decisions and perform interventions without confusion about whether they are justified in providing such interventions. Nevertheless, these concepts and distinctions become useless in limiting medical necessities. What are regarded as health and disease by practitioners and patients are different from the "health and disease" in the view of

objectivism of health. Proponents of a value-neutral theory such as Boorse (1975, 1977) refute "disease and health" in the constructive sense in this account (See §4.3).

James E. Sabin and Norman Daniels (1994), on the other hand, argue that any attempt to draw a line between treatment and enhancement is pointless, in the sense that "the treatment-enhancement distinction by itself does not specify the boundary between obligatory and nonobligatory medical services" (Daniels, 2000, p. 316). In Sabin and Daniel (1994) and Daniels (2000), several borderline cases are described to show that the distinction can be arbitrary if it is applied in practice, e.g., in insurance coverage. There are situations in which medical services only cover those with an underlying disease or condition, yet not those without an underlying condition who suffer in the same way. For instance, growth hormone treatment for children with growth hormone deficiency caused by a brain tumor is covered by health insurance, but another child of the same height would not benefit from this treatment, though he also suffers due to the preferences of society. There are other cases too, such as a man being shy because of his bipolar disorder versus a man without the condition who suffers simply due to shyness. Daniels (2000) asks:

> Does the concept of disease underlying the treatment-enhancement distinction force us to treat relevantly similar cases in dissimilar ways? Are we violating the old Aristotelian requirement that justice requires treating like cases similarly? Is dissimilar treatment unfair or unjust? (p. 312)

Sabin and Daniels contend that the distinction should be drawn in accordance with the boundary of "medical necessity", and distributive justice of medical services should be taken into consideration along with the concept of disease, since healthcare is regarded as the means for creating equal opportunity.

Yet we can ask, equal opportunity to what? Sabin and Daniels (1994) provide three models of medical necessity to illustrate the subtle differences between different goals of medical care. First, the *normal function model* claims that the main aim of medical care is to provide equal opportunities. It should thus provide the opportunities one may have had without a pathological condition, and therefore, medical necessity merely includes restoring and compensating for limited opportunity due to loss of function caused by disease and disability. The *capability model*, broadening the concept of "medical necessity", contends that medical care should aim to provide equal capacities for individuals. In Sabin and Daniels' words,

the goal of the normal function model is to help people become "normal competitors", while the capability model strives to make them "equal competitors" (Sabin & Daniels, 1994, p. 10). According to the *welfare model*, medical services should activate when one suffers from "attitudes or behavior patterns" that one did not choose to develop and is unable to overcome by oneself. That is, the suffering man with unsatisfactory height and the shy man should be offered medical services if they cannot help themselves independently.

Which model should be adopted? A useful model, in Sabin and Daniels' perspective, must be considered fair by the public and by clinicians, must be applicable in real practice, and must be affordable by society. The normal function model, in their eyes, satisfies these criteria. As Daniels (2000) puts it, "a theory of justice in general, or of justice for healthcare in particular, must combine concerns about equality with concerns about liberty, and both of these concerns must be reconciled with considerations about efficiency and the allocation of resources" (p. 317).

Sabin and Daniels' concept of enhancement based on the normal function model is similar to the idea of enhancement proposed by David DeGrazia (2005a):

> [E]nhancements are interventions to improve human form or function that *do not respond to genuine medical needs*, where the latter are defined
> 1. in terms of disease, impairment, illness, or the like,
> 2. as departures from normal (perhaps species-typical) functioning, or
> 3. by reference to prevailing medical ideology. (p. 263)

However, different from the disease-based account and the view from Engelhardt, the normal function model considers medical necessity, and the concept of treatment depends on public agreement, in which equal opportunity of individuals and distributive justice plays a relatively large part. By contrast, "enhancement" refers to medical interventions beyond medical necessity.

However, others, like LeRoy Walters and Julie Gage Palmer (1996), understand the term differently. "Enhancement", for them refers to any intervention that improves physical, intellectual, or moral traits. Nevertheless, they made a distinction between "health-related enhancement" and "non-health-related enhancement":[25]

[25] Walters and Palmer (1996) make the distinction in the chapter 4 where they discuss enhancement genetic engineering; however, in the introduction of the same book, they pointed out the difference

> In our discussion, a further distinction may become quite important, namely, the distinction between health-related physical and non-health-related physical enhancements. This distinction roughly parallels the distinction between surgery for the treatment or prevention of disease and cosmetic surgery. (Walters & Palmer, 1996, pp. 109-110)

This distinction doesn't solve the problems here, but moves the problem somewhere else. Is there a clear differentiation between health-related and non-health-related enhancement? What is health? If the notion of health relies on the "normal range", and health-related and non-health-related enhancements respectively either bring "people who fall below the normal range to achieve functioning within the normal range", or are for "persons who already functioning within the normal range" (Walters & Palmer, 1996, pp. 131-132), the reliance on statistical normality and species-typical functioning is therefore problematic. I will use the term "improvement" to refer to Walters and Palmer's idea of enhancement throughout this dissertation.

Julian Savulescu, Anders Sandberg, and Guy Kahane (2011) defend the *welfarist account* of enhancement, according to which enhancement refers to "[a]ny change in the biology or psychology of a person which increases the chances of leading a good life in the relevant set of circumstances" (p. 7). That is, any intervention is considered enhancement as long as it results in the increase of one's well-being. Unlike the disease-based account, the welfarist account is entirely normative, and relies on a conception of well-being (Earp, Sandberg, Kahane, & Savulescu, 2014). According to Savulescu, et al. (2011), what well-being is depends on the account of well-being we endorse: hedonistic, desire-fulfillment, objective list theories abound.

Like the concept adopted from Walters and Palmer (1996), and unlike others we have seen, instead of understanding "enhancement" as contrasting "treatment", the welfarist account considers treatment part of enhancement. Three subclasses of enhancement are medical treatment of diseases, increasing natural human potential, and superhuman enhancements. The third is different from the second in that it enhances one's capability beyond the range typical for the species. Savulescu et al. (2011), believe that such a concept, which reframes the current debate, allows us to reconsider what constitutes a good life (well-being) without mistakenly resisting

---

between prevention, treatment, cure of disease, on the one hand, and enhancement of capabilities or characteristics, on the other hand.

enhancement because of an overly narrow concept, which only refers to the increase of a certain functioning.

As reviewed above, approaches to enhancement can vary in different ways. These include whether a treatment-enhancement distinction is required, how the distinction is made, and how is enhancement be determined. In the next section, I will argue for the requirement of a kind of distinction.

4.2.2 The Function of the Distinction and the Health-Based Concept of Enhancement

One might question how meaningful it is to argue for a distinction between treatment and enhancement. In this section, I consider the function and the significance of a treatment-enhancement distinction. In doing so, it becomes clearer what kind of distinction we are searching for. Different from theories of Savulescu, et al. (2011) and Walters and Palmer (1996), in which a distinction between treatment and enhancement is not required, I argue that for a number of reasons, such a distinction is necessary and significant.

First, there are some conditions especially those associated with suffering that depend on different sorts of interventions, for example the infection of bacteria. Therefore, in some cases treatment and enhancement may involve dramatically different methods. However, this does not apply to all cases. What leads to the necessity of a distinction are two following reasons.

A treatment-enhancement distinction is important because there are some circumstances that should be treated differently, and these are circumstances in which one undergoes suffering. As I have introduced in §4.2.1, suffering bears a distinct moral weight that minimizing suffering is most urgent in any context. Also, because of the nature of suffering, when one undergoes suffering, one loses certain capacities, such as reasoning and planning for one's long-term goals (see §4.2.2). Then, there are pragmatic reasons: Practically, it is required for us to make such a distinction to determine, for instance the domain of medical necessity and the responsibility of a doctor.

However, there are too many (normative) issues involved if we want to determine the permissibility of an intervention, and there are individual differences and other factors involved that vary case by case. Searching for a universal criterion which directly indicates moral permissibility in all cases, may be unproductive and limit the possibilities of the discussion. Therefore, the kind of distinction I propose here is a conceptual distinction. The function of a conceptual distinction is

characterized nicely by Daniels (2000) as a "moral warning flag" (p. 320). Although the conceptual distinction does not directly imply the permissibility of interventions, and the ethical considerations involved are different from case to case, such kind of distinction shows the difference in the applicability of ethical concerns in treatment and enhancement. That is, circumstances involving treatment is much more urgent than those involving enhancement, and many ethical issues can be dismissed in the case of treatment.

4.3 The Concept of Health

How are the concepts of health and illness or disease determined? The issue of defining these concepts is related to the roles that normative and descriptive judgments play: Normative judgments distinguish good or desirable properties or ways of living from bad or undesirable ones; descriptive judgments concern the natural functioning of human or mechanisms of biological systems (D. Murphy, 2009). How much weight these two kinds of judgments carry in a theory of health characterizes the debate between objectivism and constructivism about health. Philip Kitcher (1997) has summarized the debates as follows:

> Some scholars, *objectivists about disease*, think that there are facts about the human body on which the notion of disease is founded, and that those with a clear grasp of those facts would have no trouble drawing lines, even in the challenging cases. Their opponents, *constructivists about disease*, maintain that this is an illusion, that the disputed cases reveal how the values of different social groups conflict, rather than exposing any ignorance of facts, and that agreement is sometimes even produced because of universal acceptance of a system of values. (pp. 208-209)

From this debate, theories of the nature of health can roughly be examined as a spectrum. At one end of the spectrum, strong constructivism contends that the concept of health is purely constructive, and at the other end, the concept can be found merely through a scientific theory. Strong constructivists reject a natural, objectively definable category as health or disease. For them, disease is defined purely through value-judgments: Disvalued states are first identified, and the underlying biological process is to be found (D. Murphy, 2009).

By contrast, objectivists typically focus on normal functioning and on breakdowns in functions of particular organs or systems. The functioning of

organisms sometimes departs from its "natural" or "normal" function, and this departure may be beneficial, harmless, or harmful. Harmful functioning is considered disease. A strong objectivist will claim that the determination of malfunction is an objective matter and can be determined only by science; that is, science alone can tell us what "natural" or "normal", and "harmful" mean. For them, health and disease are value-free concepts.

In §4.3, I first review different concepts of normality. Then, I look into two theories of health—the biostatistical theory from Christopher Boorse and the holistic theory of health from Lennart Nordenfelt. Discussion of these theories will form the base of my argument for a modified version of the holistic theory.

4.3.1 Normality and Normalization

We are living in a world full of norm representations. Implicitly or explicitly, we have some ideas of what the norms are. This is reflected in our behavior, values, and preferences. But the term "normal" is used in a variety of ways. Here I introduce various concepts distinguished by Murphy (1966), and discuss the interaction between normality and normalization suggested by Metzinger and Hildt (2011).

What is normal? Edmond A. Murphy (1966) has differentiated seven ways of understanding the concept; the table below comes from E. A. Murphy (1966, p. 33) and shows seven meanings that, in his opinion, are substantial in medicine, their domains of use, and suggestions for the replacements of the terms.

Based on Murphy's list, Davis and Bradley (2000) elaborate the usages of the terms:

> In medicine, *normal* can refer to a "defined standard," such as normal blood pressure, a "naturally occurring state," such as normal immunity; or simply mean "free from disease," as in a normal Pap smear. It can mean "balanced" as in a normal diet, "acceptable" as in normal behavior, or it can be used to describe a "stable physical state." In all of these meanings, the word *normal* is used to describe an "ordinary finding" or an "expected state." But medicine allows another meaning for the word *normal* that differs significantly from the ordinary. In many ways, medicine has come to understand *normal* as a "description of the ideal." (p. 8)

What I intend to show here is that there are diverse meanings of the concept, and whenever we use the term, we should specify the sense that we mean.

Table 2. *Different Conceptions of Normal.*

	Paraphrase	Domains of use	Preferable term
1	Having a Gaussian distribution	Statistics	Gaussian
2	Most representative of its class	Descriptive science (e.g., biology)	Average, median, modal
3	Commonly encountered in its class	Descriptive science	Habitual
4	Most suited to survival and reproduction	Genetics, operations research quality control, etc.	Conventional
5	Carrying no penalty	Clinical medicine	Innocuous or harmless
6	Commonly aspired to	Politics, sociology, etc.	Conventional
7	Most perfect of its class	Metaphysics, esthetics, morals, etc.	Ideal

Note: Adapted from "A scientific viewpoint on normalcy," by E. A. Murphy, 1966, *Perspectives in Biology and Medicine, 9,* p. 333-348.

A statistical approach to defining normality is based on the premise that an abnormal (state) is statistically rare: the majority of values nearest to the mean average are considered "normal", and by contrast, the minority of values farthest from the mean is "abnormal." For instance, the normal height, blood pressure, and other variables of physical or psychological conditions from textbooks or health guides are commonly derived this way, e.g., normal height is found in accordance with the range around average height.

There are different ways of determining normality (in the descriptive sense) or the reference range (in replace of the normal range). In Murphy's (1966) category, "Gaussian" and "habitual" are normality in the descriptive sense. The most common way to determine normality is to analyze data with the tool of Gaussian distribution. According to this method, the normality or the reference range is determined as the range within two standard deviations of the mean in the Gaussian distribution of the adjusted data (the range includes around 95% of the sample); whereas abnormality lies in the ranges of ~ 2.3% at either end.

However, it is important to note that the result of this approach tells us nothing about what disease is and is not. First, what disease is depends on the theory of health adopted. One of course can claim that those who fall outside the two standard deviations of the mean in the Gaussian distribution should naturally be regarded as pathological. However, an argument is required, because a condition with a value outside the reference range can be consistent with health. Second, the problem of determining disease by normality is also related to the second point: There are too many ways to analyze the variables and not every analysis will be meaningful. For instance, one may consider separating data according to sex, age, race, etc. How we approach a fruitful sense of normality depends on theoretical and empirical studies on our mind and body, and perhaps on an ethical theory about what disease and health actually are.[26] "Normalization in the normative sense" depends on the individuals' value systems and the social norms—what members constituting the society consider normal. Typical examples of social norms include appropriate dress, relationships, and manners. They are dynamic and usually culture-specific.

Engelhardt's conception of treatment (§4.2.1) relies heavily on normalization in the normative sense. This "constructive sense" of health and disease is visible when we consider the fact that there are conditions regarded as disease in one culture or era but not in another. For instance, in Chinese medicine, there is a concept of "internal heat" ("Huo-Chi") which refers to a holistic property of the human body. A high internal heat could be caused by lack of rest, tiredness or "hot" food (alcohol, ginger, litchi, etc.) Under this state, one can feel an internal heat from the body, which is often accompanied by symptoms such as a dry mouth, mouth ulcer, hemorrhoid, anxiety, etc. There are Chinese medicines that treat this kind of state, but no such concept in the western world.

One interesting point indicated by Metzinger and Hildt (2011, pp. 247-248) is the dynamic interaction between normality and normalization. They illustrate the process of normalization in the context of CE:

> Normalization is a complex sociocultural process by which certain new norms become accepted in societal practice. This process is of a more obvious political nature, because it typically involves powerful forces, like the pharmaceutical industry's continuous attempt to control the

[26] Objectivism of disease (cf. §4.3) contends that a scientific theory is sufficient for informing us what disease and health are.

> implementation and the marketing of new enhancement technologies in society, for example, by changing the medical taxonomy or inventing new diseases and theoretical entities. The scientific process—say, of optimizing textbook definitions, predictions, and therapeutic success—is highly political as well. It attempts to firmly anchor theoretical entities like "normal mental functioning" or "normal age-related cognitive decline" in empirical data, but it is also driven by individual career interests, influenced by funding agencies, media coverage, and so on. In addition, it may well be that concepts like that of a "cognitively healthy individual" will always possess an *irreducibly* normative component. (pp. 347–348)

Such a complex process or normalization results in a change in the criterion of normality (an increase in the prevalence rate). On the other hand, the increase of the prevalence rate will in turn modify the mainstream conception and attitudes toward something (e.g., undergoing CE).

The task here is not to determine which concept of normal should be adopted, but to ask what it could mean when the term "normal" is used with a view to further approaching a more precise concept of health and enhancement. It is also important to note that though normality and normalization are two independent concepts, they interact and refer to dynamically changing phenomena.

4.3.2 Christopher Boorse on Biostatistical Theory

Christopher Boorse (1977) clearly states his objectivist claim: "On our view disease judgments are value-neutral. […] If diseases are deviations from the species biological design, their recognition is a matter of natural science, not evaluative decision" (pp. 542-543). According to Boorse, the concepts of health and disease can be determined by normality in the descriptive sense. Health is normal functioning. The concept of normal here is not Gaussian, as in E. A. Murphy's classification of the use of normal mentioned in §4.3.1 (which Boorse refers to as statistical normality of clinically observable variables), but the statistical normality of internal physiological functions. The former is merely an imperfect guide to the latter, whereas the pathological is determined by statistical species-subnormality of biological function relative to sex and age (1997, p. 32).

With reference to Boorse's "Biostatistical Theory" (BST), "[a] *disease* is a type of internal state which is either an impairment of normal functional ability, i.e., a reduction of one or more functional abilities below typical efficiency, or a

limitation on functional ability caused by environmental agents" (Boorse, 1997, p. 7). Health is identical to the absence of disease: Health is normal functioning, which is, relative to sex, age, and race, the *statistically typical* contribution of all the organism's parts and processes to the organism's goals of its survival and reproduction (Boorse, 1977, p. 555), while disease is deviation from this natural functioning (Boorse, 1975, p. 59).

In a value-free theory of health and disease, two important notions are "function" (in teleological sense) and "normality" (in a statistical sense). The notion of function is of "a contribution to a goal". The structure of the organism is constituted by "a mean-end hierarchy with goal directedness at each level" (Boorse, 1977, p. 556): The parts or processes of the lower-level, by fulfilling their own goal-directed tasks, contribute to higher-level goals, and the goal of the whole organism is individual survival and reproduction. With regard to "normal" functioning, this is "natural" functioning, which means "the performance by each internal part of all its statistically typical functions with at least statistically typical efficiency" (Boorse, 1977, p. 558).

Boorse's objective value-free account defines health and disease by shifting the burden to the notion of normal functioning, and this is grounded in the goals of survival and reproduction. This leads to some controversial cases. One such controversial case is homosexuality. Another similar case is childfree people or voluntary childlessness: people who are fertile but do not intend to have children. To human beings, survival, reproduction, and related goals are not necessarily central purposes (see §6.3.2). Some kinds of human behaviors, such as altruism, as well as human history, have shown us that there are "greater goals" (e.g., science, art, and human rights) considered by human beings, which cannot be easily explained by Darwin's theory of evolution.

A weaker version of objectivism does not preclude the involvement of normative judgments whose role is to categorize natural kinds. For weak objectivists, the concept of disease, unlike concepts such as "dandelion" or "monkfish", is close to the idea of "weed" or "vermin" (D. Murphy, 2009). The former is regarded as naturally existing category of the world and can be investigated by science, while the latter depends on human interests and is value-laden.

4.3.3 Lennart Nordenfelt on the Holistic Theory of Health

The rival theory against BST is what Lennart Nordenfelt (2007) has termed the "Holistic Theory of Health" (HTH). In addition to the goals of reproduction and survival as the criteria of health in BST, HTH regards individuals as social agents and emphasizes the "quality of life" of the individual: "[A] person can be ill, not only if the probability of the person's survival has been lowered but also if he or she does not feel well or has become disabled in relation to some goal other than survival" (Nordenfelt, 2007, p. 6).

HTH is different from BST in the following respects. First, it emphasizes the relationship between an individual and the social network in which she lives. Second, while BST relies on statistical considerations, HTH is based on an idea of human welfare. Third, HTH contrasts "health" with "illness" rather than with "disease". The difference between these two notions is addressed in §4.4.1.

According to HTH, there are two features of the concept of health: (1) Health and illness are respectively accompanied by "the feeling of wellbeing" and "the feeling of suffering"; (2) the concept of health is characterized by the ability to realize a certain kind of goal in life. Nordenfelt's (1993) Welfare-theory of health defines health and illness as follows:

> P is completely healthy, if and only if P has the ability, given standard circumstances, to realise all his or her vital goals.
> P is unhealthy (or ill) to some degree, if and only if P, given standard circumstances, cannot realise all his vital goals or can only partly realise them. (p. 280)

Nordenfelt contends that a vital goal includes not only survival but also a minimally decent life: life without disabling pain, with realization of the most important projects of the individual. The "vital goals" Nordenfelt refers to are "the states of affairs which are necessary and jointly sufficient for his or her minimal long-term happiness" (Nordenfelt, 2001, p. 67). "Happiness" as a "want-related" concept is "an equilibrium between the subject's wants and the world as he or she finds it to be" (p. 68). But what is "minimal" happiness?

> When the life-situation of people is at a level where they hesitate as to whether they are on the whole happy or unhappy, then they are, I would say, on the minimal level of happiness. The minimal level is such that below it

> the person is unhappy, while above it the person is happy, not necessarily completely happy, except for the limiting case. (Nordenfelt, 2001, p. 105)

Overall, to be healthy is not identical to being minimally happy in the long run, but a person is healthy if and only if she has the ability, in standard circumstances, to realize the goals which are necessary and sufficiently for her minimal happiness.

4.4 The Phenomenological Account of Health and Enhancement

4.4.1 The Reverse Theory of Disease and Illness

A phenomenological account of health (PAH) distinguishes health from ill-health by the possession of the ability to escape from suffering. Below, a distinction between disease and illness is introduced, from Andrew Twaddle's triad, where disease is treated as a derivative concept of illness. Then, inspired by negative utilitarianism, an absence of suffering is used to characterize the concept of health. Based on these ideas, the PAH is proposed and examined.

Andrew Twaddle first introduces and applies the triad of disease, illness, and sickness in his doctoral dissertation of 1967, and the distinction has been used in medical related disciplines ever since (as cited in Hofmann, 2002). These three terms respectively refer to objective, subjective, and social aspects of ill-health (p. 651). First, disease refers to the physiological event in a health problem: It is the physiological or neurobiological malfunction that results in an actual or potential reduction of physical or mental capacities or a reduced life expectancy. Disease can only be measured by objective means. The subjective aspect of ill-health is delineated by illness. Illness is the subjective undesirable feeling directly perceived by the individual, such as pain, weakness, dizziness. Last, concerning the social aspect, the concept of sickness is defined by the members in a society: "Sickness is a social identity. It is the poor health or the health problem(s) of an individual defined by others with reference to the social activity that individual" (Twaddle, 1994, cited in Hofmann, 2002, p. 651).

Here disease, illness, and sickness are treated as three independent concepts in contrast to health:

> The paradigm case in health care is when a person feels *ill*, the medical profession is able to detect *disease*, and society attributes to him the status *sick. Illness* explains the person's situation to himself, *disease* permits medical attention, and *sickness* frees him from ordinary duties of work and

> gives him the right to economic assistance […]. (Hofmann, 2002, pp. 657-658)

Not only do these three concepts reflect three perspectives—biological, phenomenal, and social—of a health-related event, but also do they have dynamic interactions with each other. Boorse (1975, p. 61) has made a distinction between disease and illness in the past, but illness plays no role in his Bio Statistical Theory, for the theory relies on biostatistically normal functioning without reference to the subjective aspect of that functioning. Conversely, Nordenfelt (2007, p. 7) considers illness—the subjective aspect of ill-health—to be the primary notion.

According to the Reverse Theory of Disease and Illness (Canguilhem, 1978; Fulford, 1989) Nordenfelt (2007) endorses:

> The primary focus of attention is thus the illness—the problem as perceived normally by the subject. From the concept of illness we can derive the concept of disease, i.e. the internal state which causes (or tend to cause) the illness. But observe here how the diseases are identified. They are identified on the basis of an illness-recognition. A discovery of the disease presupposes the occurrence of an illness. (Nordenfelt, 2007, p. 8)

I agree with Nordenfelt's claim that the subject's phenomenal experience should be given primacy. Illness characterizes the negative phenomenal state of the subject, while disease refers to the biological mechanism corresponding to illness and potential illness. The distinction and order of illness and disease will be the first building block of the phenomenological account.

Although I support the primacy of illness, I disagree with Nordenfelt's view that "minimal happiness" forms the division between health and ill-health. Minimal happiness, as Nordenfelt states, is the equilibrium between the subject's desire and satisfaction. For two reasons, I propose "minimal suffering" as the borderline, rather than minimal happiness. First, we seem to share the ethical intuition that unnecessary suffering should be avoided (§4.1.1). Based on the negative utilitarianism, we ought to minimize overall suffering. By using "minimal suffering" to distinguish health, it allows the distinction between health and ill-health to indicate the urgency of two different kinds of states. Second, suffering has a biological basis in urging the organism to escape from a harmful situation (§4.1.2). When one is under suffering, one's mental resources will be occupied to deal with

the source of the current danger. In §6.3.2, I will show that this can influence one's capacity of decision-making, and thus one's autonomy will be treated differently.

4.4.2 The Phenomenological Account of Health

Disagreeing with Boorse's (1975, 1977, 1997) Biostatisical theory, which relies on the problematic concept of biostatistical functioning, I side with Nordenfelt's (1993, 2001, 2007) stress on the subject's phenomenal state: The first-person experience of the subject takes primacy over the Darwinian goals proposed by Boorse. However, by contrast to Nordenfelt's Welfare-theory of health, I propose minimal suffering as opposed to minimal happiness.

According to the PAH,

> a subject S is healthy if and only if under standard circumstance, S has the ability to avoid or escape from occurrent or potential suffering. S is unhealthy if and only if under standard circumstances, S is currently undergoing suffering and is not able to escape from the situation, or is prone to potential suffering.

"Standard circumstance", following József Kovács' (1998) analysis of Nordenfelt's theory, are the typical environments in which most individuals in a group live. This represents the "external factors with which practically every member of a society is confronted" (p. 36) including norms, institutions, preferences, ideologies.

One may question whether the PAH is a reductionist theory. This question can be understood in two senses. First, in light of the debate between objectivism and constructivism, does the account presuppose un-naturalized moral judgments? To begin with, the PAH is certainly not value-free. It involves a normative judgment of the moral significance of suffering. However, this dissertation only adopts the negative utilitarianism as a practical theory; it does not imply the endorsement of either cognitivism or non-cognitivism, according to which moral judgments can or cannot be attributed truth conditions (van Roojen, 2012). The other way of understanding the question concerns emphasis on the phenomenal aspect of the subject. This does not imply the ontology of a mental state or a denial of materialism. A physicalist or functionalist theory of mind may argue that suffering can be reduced to brain regions or functional properties, but the PAH is independent from debate of the mind-body problem.

It is worth mentioning that, parallel to the distinction of normality in the descriptive sense and normalization in the normative sense, and to their interaction (see §4.2), apart from the concepts of health, illness and disease characterized in the phenomenological account, there are the other senses of health and ill-health which are based on each individual's value system and the social norm she is embedded in. This "normative sense of health and ill-health" influences the concepts of health, illness, and disease, which rely on the existence or absence of suffering, because suffering becomes the result of a variety types of social coercion. Likewise, the former can be affected by the latter: As we know more about our suffering, our understanding of health, illness, and disease evolve.

4.4.3 The Phenomenological Account of Enhancement

Here, I propose the PAE, based on the HAE, according to which the distinction between treatment and enhancement relies on the distinction between health and illness (§4.2.2). I also examine the PAH, according to which the phenomenological concept of heath refers to the state in which one is free from occurrent or potential suffering (§4.4.2). According to PAE, the demarcation between treatment and enhancement depends on whether there is current or potential suffering resulting from the dysfunction the intervention addresses:

> *Treatments* are, under standard circumstances, interventions that address a malfunction that results in an individual's suffering or potential suffering, and which the subject is not able to independently avoid or escape from.
> *Enhancements* are, under standard circumstances and without any unwilling suffering or potential suffering resulting from a target function, interventions that aim to manipulate the target function based on the subject's interests.

Such an account distinguishes the concepts based on negative phenomenology, because of the moral significance of suffering (see §4.1).

First, to utilize the phenomenological concept of enhancement in the context of CE that aims at different kinds of function (e.g., memory enhancement targeting memory function), it is required to examine the relation between suffering and the malfunction of the cognitive function in question. One may enhance one cognitive function while at the same time suffer from another cognitive dysfunction. It is therefore important to note how suffering results from different kinds of cognitive dysfunctions.

Second, I define enhancements as “manipulating” rather than “optimizing” functions, because the optimization of one functional capacity does not necessarily imply a better cognitive function or a reduction of suffering. The concepts of enhancement and improvement are normative concepts: They involve the issue of what is *good* or *better*. By definition, the state of a being after enhancement or improvement is better than the previous state. The question is whether and how we can we find a universal criterion of what is better (e.g., a better physical state or cognitive function).

Moreover, the concept of enhancement is commonly used in the context of enhancement of particular human traits, such as a cognitive or physical capacity. Nevertheless, how do we determine which kind of capacity is better? For instance, is memory capacity strengthened by intervention really better? Some cases suggest that more is not better. For example, the Russian journalist, S. V. Shereshevskii, who seems to have unlimited memory capacity, suffers from several difficulties, such as the problem of understanding abstract concepts. A. R. Luria records how he was unable to grasp the concepts “infinity” or “nothing” in *The Mind of a Mnemonist*:

> . . . *Infinity*—that means what has always been. But what came before this? What is to follow? No, it's impossible to see this . . . In order for me to grasp the meaning of a thing, I have to see it . . . Take the word *nothing*. I read it and thought it must be very profound. I thought it would be better to call *nothing* something . . . for I see this *nothing* and it is something . . . If I'm to understand any meaning that is fairly deep, I have to get an image of it right away. So I turned to my wife and asked her what *nothing* meant. But it was so clear to her that she simply said: *"Nothing* means there is nothing." I understood it differently. I saw this *nothing* and felt she must be wrong. The logic we use, for example. It's been worked out on the basis of years of experience. I can see how it has developed, and what it means to me is that one has to rely on his own sensations of things. If *nothing* can appear to a person, that means it is something. That's where the trouble comes in . . . (Luria, 1968, p. 131)

This case shows that maximizing memory capacities does not directly result in a better memory, in the sense that it does not necessarily promote one’s well-being.

Furthermore, what is considered better differs from person to person. Accordingly, I propose that what kind of intervention is to be enhancement depends on each subject's interests—whether the subject identifies and prefers the expected result and the intervention itself. More will be elaborated in §6.3.2.

4.5 Summary

This dissertation is based on a practical and minimal normative theory—negative utilitarianism. Negative utilitarianism and suffering are first introduced.

- *The principle of negative utilitarianism is suffering-minimization: We ought to minimize overall suffering (§4.1.1).*
- *Suffering is a phenomenal event available only to beings with the capacity to feel. It has the function of warning and compelling the subject at a conscious level to take action to free itself from an unpleasant situation. Suffering forces one's mental resources to tackle the problem that suffering results from (§4.1.2).*

Then, different definitions of enhancement are reviewed. I propose a modified version of the disease-based account—the health-based account of enhancement (HAE).

- *According to HAE, enhancements are distinguished from treatments by what they respond to: Enhancements aim at healthy states, whereas treatments aim at unhealthy states (§4.2.1).*
- *The distinction between treatment and enhancement functions as a "moral warning flag", which highlights the different kinds of moral issues that must be taken into account (§4.2.2).*

As for the concept of health, I dispute Boorse's (1975, 1977, 1997) objectivist account of health and by modifying Nordenfelt's (1993, 2001, 2007) welfare-theory of health (§4.3), I propose a phenomenological account of health (PAH).

- *Boorse's objectivist account is disputed, because, based on negative utilitarianism, one's subjective experience (free from suffering) should be taken into account, instead of normal functionality. Unlike the welfare-theory, I take existence of suffering as the defining criterion for*

"illness" in contrast to "health", instead of Nordenfelt's proposal of minimal happiness.

- *According to PAH, one is healthy if and only if, under standard circumstances, one has the ability to avoid or escape from occurrent or potential suffering; one is unhealthy if and only if, under standard circumstances, one is currently undergoing suffering and is not able to escape from the situation, or is prone to potential suffering (§4.4.2).*

Derived from the HAE and the PAH, the phenomenological account of enhancement (PAE) is developed.

- *According to PAE, treatments are, under standard circumstances, interventions that address a malfunction that results in an individual's suffering or potential suffering, and which the subject is not able to independently avoid or escape from; enhancements are, under standard circumstances and without any unwilling suffering or potential suffering resulting from a target function, interventions that aim to manipulate the target function based on the subject's interests.*

Chapter 5

Memory Interventions:
The Current Situation

5.0 Introduction
5.1 Interventions at the Molecular Level
5.1.1 Neuroplasticity and Memory
5.1.2 Mechanisms of LTP and LTD
5.1.3 Pharmaceutical Interventions
5.1.4 Genetic Engineering
5.1.5 Herbs and Nutrition
5.2 Brain Stimulations
5.2.1 Deep Brain Stimulations
5.2.2 Transcranial Stimulations
5.3 Physical and Mental Exercises
5.4.1 Sleep
5.4.2 Mnemonics
5.4 The Classification of Memory Interventions
5.5 Summary

5.0 Introduction

Thanks to advances in medicine, we have been able to extend our lifespan so that it is much longer than in the past. The extension is due to fewer lethal physical diseases and greater capacity to maintain physical conditions and to constrain such diseases. In contrast, we do not yet have as much control over our mental condition. The situation whereby our "physical life" outlives our "mental life" is reflected in the increasing number of age-related neurodegenerative disorder (e.g., dementia). There has been more and more emphasis on the importance of high physical and mental quality of life, as well as an increasing number of studies on psychological disorders and treatments.

In understanding human cognition, memory has played a central role (see §2); however, it is a cognitive function that declines comparatively early and significantly. Increasing levels of dementia and mild-cognitive impairment have

motivated pharmaceutical companies to devote themselves to developing new memory-intervening substances that prevent and treat memory-related disorders. This chapter will review different ways of intervening in memory function, the plausibility of applications of such treatments on healthy individuals, as well as their classification.

First, after briefly sketching out the molecular mechanisms of memory formation and consolidation, pharmaceutical interventions and genetic engineering that aim to improve memory are introduced (§5.1). Then, herbs such as *Ginkgo biloba* and nutrition supplements, such as glucose, are reviewed (§5.2). Brain stimulation, including invasive and non-invasive brain stimulation, is shown to have a positive effect on memory (§5.3). Last, other physical and mental exercises are discussed (§5.4).

As we have seen in §2, memory processes are constituted by several stages, including encoding, persistence, retrieval, consolidation, and reconsolidation. Memory enhancers can intervene at each stage of this process. In addition, performance in memory tests can also be improved indirectly; for instance, enhancing attention may contribute to the probability of the information being encoded and will therefore result in the improvement of memory. This chapter mainly focuses on interventions that directly improve memory.

5.1 Interventions at the Molecular Level

5.1.1 Neuroplasticity and Memory

Formation of new memory involves two phases, including acquisition (forming short-term memory) and consolidation (from short-term memory to long-term memory. Short-term memory (STM) decays rapidly (within minutes) and is vulnerable to various forms of disruption, whereas long-term memory (LTM) lasts for days—or even a lifetime—and is more resistant to disruption. (The content of LTM can still be disrupted through the process of reconsolidation.)

Enabling the formation of STM and LTM is the plasticity of the nervous system. Neuroplasticity is the capacity of a nervous system to alter its structure, organization, and function to adapt to changing requirements (Mateer & Bogod, 2003, p. 168; Nava & Röder, 2011, p. 177). The idea of plasticity was first proposed by Ramón y Cajal as the synaptic plasticity hypothesis in order to solve a paradox (cited in L. R. Squire & Kandel, 2009, p. 38): If learning and memory result from a change in nerve cells, but the connections of neurons are for the most part already

established, what kinds of changes lead to plasticity? (Only recently have scientists discovered neurogenesis (i.e., the generation of neurons) of the adult mammalian brain (Altman, 1962) and human brain (Eriksson et al., 1998).) Cajal hypothesized that the strength of synaptic connections is plastic and modifiable, and learning and memory are respectively carried out by the alteration and persistence of strength of synaptic connections. This hypothesis was later examined by Kandel and his colleagues in gill-withdrawal reflex in *Aplysia* (cited in L. R. Squire & Kandel, 2009, p. 42).

Nowadays, neuroscientists are able to study exactly what happens at the synaptic level. The property of neuroplasticity can be understood through synaptic phenomena: long-term potentiation (LTP) and long-term depression (LTD). LTP and LTD are processes that respectively increase and decrease the tendency for a signal to pass from one neuron to the next within a connection, and are each triggered by high- or low-frequency stimulation. LTP can be seen in both processes of acquisition (early LTP) and consolidation (late LTP). Nevertheless, different from formation of STM, consolidation of LTM requires protein synthesis and anatomical change of synapses (Tully, Bourtchouladze, Scott, & Tallman, 2003, p. 268).

It is important to note that LTP at one synaptic connection does not imply formation or consolidation of memory. The neurons in our brain are interconnected in extremely complex ways. On the one hand, a piece of memory is realized by several different groups of neurons and the synaptic connections between them; on the other hand, a synaptic connection and a neuron contribute to the realization of different memory retrievals and even a variety of cognitive functions.

Contrary to LTP, which has been well investigated and proven to play an important role in learning and memory, the function of LTD is not yet clear. Though LTD can be found in several regions in the central nervous system, scientists still do not know if it is only engaged in forgetting or it has a function in normal learning and memory mechanism. It has been speculated that LTD in the hippocampus has the function of enabling replacement of one memory representation with another (Diamond, Park, Campbell, & Woodson, 2005). It happens when the organism is in a novel environment or an urgent situation (e.g., under exposure to predators or undergoing a painful experience) and involves an interaction between the hippocampus and the amygdala. It is adaptive, as LTD allows the new information with survival value to take priority in consolidation and retrieval.

Although facilitation of LTP and LTD in one synaptic connection does not directly lead to enhancement of consolidation of a memory, improvement in neuroplasticity will result in improvement in learning and memory. Therefore, neurobiological research into new drugs for memory improvement has focused on neuroplasticity and the mechanism of LTP and LTD (Tully et al., 2003, p. 267). The plastic property of the nervous system can be realized by a variety of mechanisms: (1) by modifying the efficacy of synaptic transmission at preexisting synapses (changing electrophysiological properties, receptor number or sensitivity); (2) by creating new synaptic connections or discarding existing ones; (3) by modulating the excitability of neurons; or (4) by forming new nerves or glial cells (Kays, Hurley, & Taber, 2012, p. 119; Malenka, 2002, p. 147).

5.1.2 The Mechanisms of LTP and LTD

The mechanisms of early and late LTP are characterized in figure 4. (The blue arrows indicate the pathway of early LTP, and the red ones represent late LTP.) Squire and Kandel (2009, p. 166) provided a complete review of the mechanisms involved in LTP. There are two kinds of glutamate receptors on the dendrite of a post-synaptic neuron: α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors (AMPARs) and N-methyl D-aspartate receptors (NMDARs). Both can be bound and activated by glutamate, which is one of the main neurotransmitters in modifiable synapse in the brain and spinal cord. During weak electrical stimulation, only the former will permit the flow of sodium ions (Na^+) and potassium ions (K^+), which result in excitatory postsynaptic potential in the post-synaptic neuron; by contrast, the channel of the latter is blocked by magnesium ions (Mg^{2+}), so that no ions are able to pass through.

When a strong electrical simulation with sufficient strength and frequency appears, the depolarization regulated by AMPARs expels the blockage from the channels of NMDARs, and thus not only Na^+ and K^+ are allowed to move across the membrane and contribute to depolarization, but calcium ions (Ca^{2+}) can also enter the post-synaptic neuron to start the subsequent intracellular signaling cascade. The Ca^{2+} that entered the post-synaptic neuron binds with and activates Ca2+/calmodulin-dependent protein kinase II (CaMKII). The latter not only (1) leads to the insertion of more AMPARs and thereby increases the sensitivity of postsynaptic neuron to glutamate but also (2) activates the enzymes that generate retrograde signals, and which feedback on the presynaptic neuron to enhance the release of transmitters. This is the mechanism for short-term plasticity.

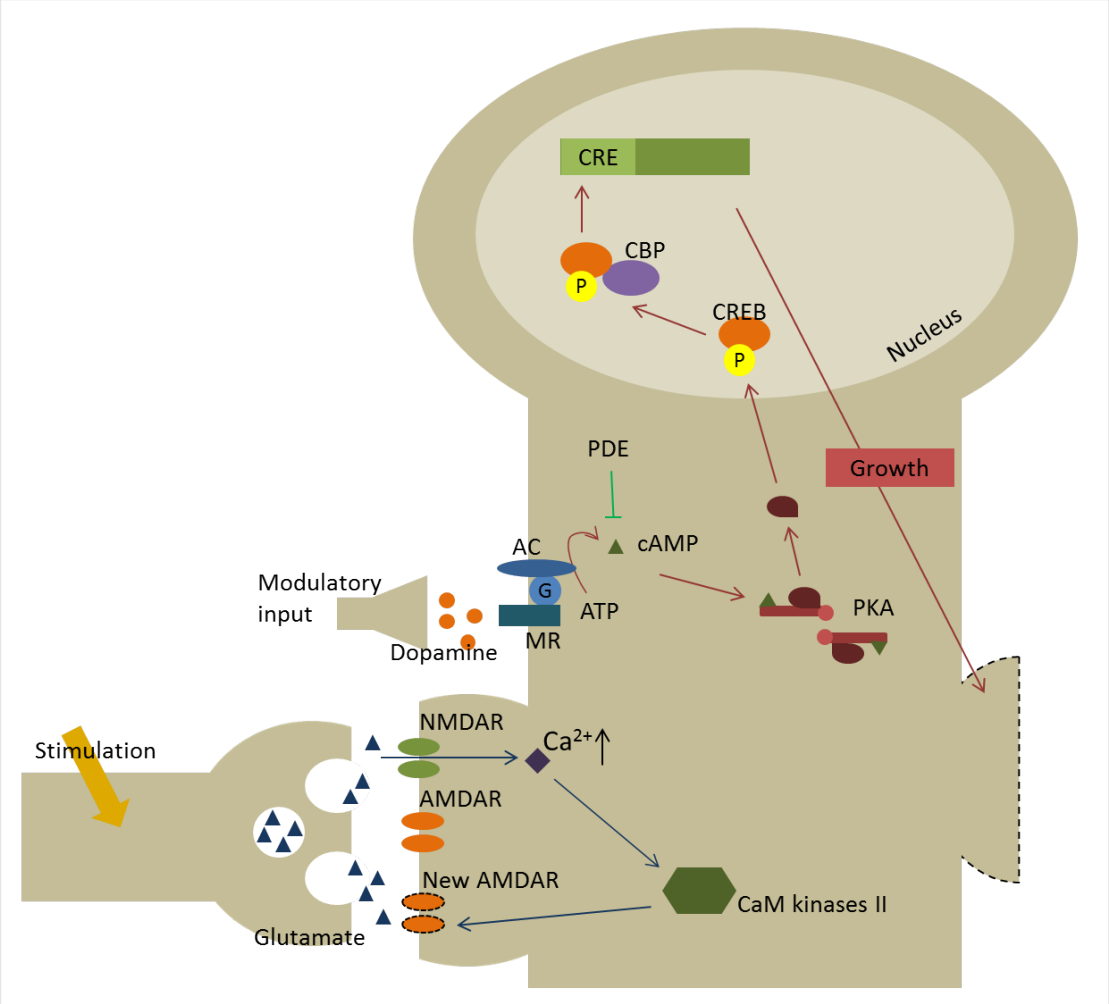

Figure 4. The molecular mechanisms of early and late long-term potentiation (LTP). The pathway of early LTP is shown with blue arrows, and the pathway of late LTP with red. AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor; NMDAR, N-methyl D-aspartate receptor; Ca^{2+}, calcium ions; CaMKII, Ca2+/calmodulin-dependent protein kinases II; MR, metabotropic receptor; G, G protein; AC, adenylate cyclase; ATP, Adenosine-5'-triphosphate; cAMP, cyclic adenosine mono-phosphate; PDE, phosphodiesterase; PKA, protein kinase A; CREB, cAMP response element-binding; p, phosphorylated; CBP, CREB-binding protein; CRE, cAMP-responsive element. Figure modified, from Stern & Alberini (2013, Figure 1) and Squire & Kandel (2009, p. 166).

Repeated training activates a dopaminergic modulatory input, which in turn activates adenylate cyclase (AC). AC catalyzes Adenosine-5'-triphosphate (ATP) to cyclic adenosine mono-phosphate (cAMP). Protein kinase A (PKA), catalyzed by cAMP, releases its catalytic subunits, which then (1) phosphorylates AMPAR to make it remain open longer and increase its conductance to ions and (2) moves to the nucleus to activate cAMP response element-binding protein (CREB). Activated (phosphorylated) CREB binds to the cAMP-responsive element (CRE) region on the DNA and recruits its transcription co-activator CREB-binding protein (CBP). This results in creation of new AMPA receptors and growth of new dendrites via gene transcription and protein synthesis.

By contrast to LTP, LTD, which weakens the strength of a synapse, is regarded as a complementary process. It occurs when the pre-synaptic neuron is active at low frequencies (1-5 Hz) without strong depolarization of the post-synaptic neuron. The enzyme, phosphatase, which is inhibited by Inhibitor 1 in LTP, is activated in LTD to dephosphorylate AMPARs phosphorylated by CaM kinase II and PKA. Besides, AMPARs will be removed from the postsynaptic membrane and placed in reserve. Both the decreased effect of AMPARs, owing to dephosphorylation, and the decreased number of receptors result in a lower level of depolarization and normal level of strength of the synapse. It is believed that this return to normal level of strength enables the storage of new information.

5.1.3 Pharmaceutical Interventions

Pharmacological studies have aimed to enhance LTP by manipulating the CREB pathway. First, at the postsynaptic site, the functions of NMDARs and AMPARs can be enhanced. NMDAR can be enhanced through phosphorylation of CaMKII (Lee & Silva, 2009, p. 129). Memantine, a kind of NMDAR antagonist, can block the activity of the neurotransmitter glutamate, for hyperactivity of the glutamatergic system leads to neuronal damage or death (Parsons, Stöffler, & Danysz, 2007). Memantine has been approved by the European Medicines Agency (EMA) and Food and Drug Administration (FDA) for treating moderate-to-severe Alzheimer's disease. However, no study of memantine on normal subjects has yet been undertaken (Dresler et al., 2013). It is speculated that memantine is only effective when the homeostasis of the glutamatergic system is impaired, and thus no effect would be expected in healthy subjects.

As for AMPARs, Gary Lynch, at the University of California, Irvine, is a pioneer in developing AMPA receptor-based memory drugs. Ampakine, a

structurally diverse family of molecules that positively modulates AMPA-type glutamate receptors and facilitates fast, excitatory transmission throughout the brain, is one of his research targets. In animal studies, it has been shown to accelerate encoding of LTM in young rats and reverse LTP deficits that appear in middle-aged rats (Lynch & Gall, 2006, pp. 558-559). In humans, small-to-moderate improvements in encoding both STM and LTM were found in young adults, and significant results were discovered when tested in 65–75 year-old healthy subjects (not symptomatic of any neuropsychiatric disturbances, including mild cognitive impairment) (Lynch, 2002, p. 1037). Moderate-to-large improvement of memory is also seen in schizophrenic patients (Lynch, 2004, p. 9). Cortex Pharmaceuticals Inc., headed by Lynch and Gary Rogers, has been studying the potential of a type of amapkine "CX-717" as a treatment for Alzheimer's disease, and recently a supplier of memory products reported that a new CX-class drug, "CX-1739", was recently proposed to be five times more potent than its predecessor (PeakNootropics, 2013, Febuary 3). Other pharmaceutical companies, such as Lilly, Organon, and Servier, are also in the race to produce AMPAR-based memory enhancers (Lynch, 2002; Pogačić Kramp & Herrling, 2011).

In addition, protein kinase C isoform M Zeta (PKMζ), which can increase the number of postsynaptic AMPARs by regulating the trafficking of receptors, is another target for modulating AMPARs (Migues et al., 2010). In studies on rats, enhancing PKMζ resulted in enhanced LTM (Stern & Alberini, 2013, p. 43). Though the underlying mechanisms are still under investigation, it is shown that PKMζ is required for maintenance of LTM even long after consolidation. It is speculated that LTM can be erased rapidly after local application of PKMζ inhibitors, and this has been proved in rats (Shema, Sacktor, & Dudai, 2007).

Second, LTP can be manipulated through postsynaptic transcription factors in the nucleus. CREB can be regulated through the actions of several kinases and phosphatases and through genetic manipulation. A very detailed overview of signaling pathways of CREB can be seen in Alberini (2009, Fig. 3). Genetic manipulations that result in decreased CREB transcription in mice lead to deficits in both LTP and long-term memory; overexpression of CREB resulted in memory improvement.

Eric Kandel, a neurobiologist at Columbia University and a co-recipient of the Nobel Prize in Physiology or Medicine in 2000 for work on the biochemistry of neuron signaling, has devoted himself to the study of how memories are formed and stored at the molecular level and to forming a model of long-term memory

consolidation. Memory Pharmaceuticals Corp., the company run by Kandel and Walter Gilbert, the Harward biochemist and a Nobel Prize winner in chemistry in 1980, has designed two drugs "MEM1414" and "MEM1917" to improve memory by sustaining cAMP levels (Marshall, 2004; Pogačić Kramp & Herrling, 2011). These two drugs are phosphodiesterase-4 (PDE4) inhibitors. By inhibiting phosphodiesterase, which hydrolyzes cAMP, they modulate CREB levels. Rolipram used to be prescribed for depression, but is also a DPE4 inhibitor (cited in Stern & Alberini, 2013, p. 42). Other than Memory Pharmaceutical Corp, Sanofi-Aventis, ExonHit Therapeutics SA, and Helicon Therapeutics are also trying to develop PDE4 inhibitors. Another way to increase cAMP is to increase the sensitivity of Adenylyl cyclase (AC). Forskolin is a small molecule agonist of AC that binds to the active site of the enzyme and stimulates the synthesis of cAMP (Tully et al., 2003, pp. 270-271).

A third way of regulating LTP is through the regulation of neuromodulators. Dopamine, which is insufficient in Parkinson's disease, is important for working memory and plays a role in LTM. In prefrontal cortex (PFC) dependent tasks, application of dopamine receptor antagonists improves performance. However, this kind of memory improvement can be seen only in aged animals, rather than young ones (Stern & Alberini, 2013, p. 45). As shown in Figure 4, dopamine binding with dopamine receptors, which is coupled with AC, and the activation of AC, leads to formation of cAMP. This starts the CREB pathway for late LTP described earlier (Jay, 2003). Dopamine agonists may be regarded as a treatment for patients with low dopamine levels, but studies have not yet shown that it has any significant influence in healthy individuals.

Other neuromodulators are involved in memory formation. For instance, stress hormones corticosteroids improve memories involving emotional arousal, and noradrenaline, which contributes to CREB pathways through activation of beta-adrenergic receptors, can improve memory retention if administrated after training (Stern & Alberini, 2013, p. 45). Nevertheless, no studies show significant improvement of memory in healthy human beings. Furthermore, people are interested in knowing whether stimulant drugs for attention deficit hyperactive disorder (ADHD), such as methylpehidate (Ritalin®) and modafinil (Provigil®), have the effect of enhancing memory. No consistent and strong evidence has shown that they have any effect on memory (Franke & Lieb, 2010, Table 1).

Acetylcholine (Ach), which is required for memory formation, is another important kind of neuromodulator. Acetylcholinesterase inhibitors (AChEIs), which

inhibit acetylcholinesterase, the enzyme for degrading ACh, including donepezil (Aricept®), rivastigmine (Exelon®) and galantamine (Reminyl®), and are one way of treating mild to moderate Alzheimer's disease (Stern & Alberini, 2013, p. 46). Among these drugs, only donepezil has been studied in both healthy young and old subjects. In Dresler, et al. (2013), six trials of donepezil were reviewed: verbal memory was shown to improve; positive and negative effects on episodic memory were seen respectively in two different trials; another study demonstrated impairment of working memory in older adults (p. 531). There is not enough data or consistent evidence to show that donepezil has potential as a memory enhancer.

5.1.4 Genetic Engineering

Since the transition from short-term to long-term memory requires gene transcription and protein synthesis, genetic engineering can be another way of improving memory. In order to improve memory storage, gene activation is involved in strengthening synaptic connectivity. Many transgenic and KO studies in mice have revealed a large number of mutations that seem to enhance learning and memory, and have also identified memory-related genes. An overview of these genes can be seen in Lee and Silva (2009, Table 1). Most of the manipulation of these genes aims at intervening in the CREB pathway, described earlier.

Li-Huei Tsai, at the Massachusetts Institute of Technology, has focused on another link, rather than CREB pathway. Tsai and her colleagues have shown that the memory of mice can be improved by unwinding DNA (Guan et al., 2009; Howard Hughes Medical Institute, 2009, May 7). A kind of enzyme, such as histone deacetylase (HDAC) proteins, that keeps DNA inside neurons tightly coiled and unable to "relax" for expression by removing acetyl groups, hampers learning and memory. Thus, compounds that block the activity of these HDAC proteins may result in memory improvement. Tsai's lab has shown that inhibiting HDAC proteins improved memory in mice with gene mutations related to Alzheimer's disease. They have tested several different drugs that inhibit HDAC activity in mice, and some of these drugs have been evaluated as potential therapies for memory loss associated with Alzheimer's disease. Tsai plans to test HDAC-blockers in mouse models of autism and schizophrenia. She acknowledged that it would take a long time for these HDAC-blockers to be tested on humans. In fact, the difficulty of predicting the potential on human models through translational animal researches may have generally been underestimated (Hackam, 2007).

5.1.5 Herbs and Nutrition

One of the most well-known herbs for memory enhancement is *Ginkgo biloba*. Whether this herb can improve memory is controversial. In dementia, Tchantchou et al. (2007) found that a Ginkgo extract, EGb 761, which has an effect on CREB, can significantly increase cell proliferation in the hippocampus of mice with the human gene for Alzheimer's disease. The hippocampus is the part of the brain mainly affected by Alzheimer's disease. The Ginkgo extract may have therapeutic potential for prevention and improvement of Alzheimer's disease. However, the benefit of *Ginkgo biloba* to human beings has been controversial. Results of studies and analysis of its effect in patients with dementia or cognitive impairment have been inconsistent (Birks & Grimley Evans, 2009). As for healthy individuals, a review on studies of cognitive enhancement has shown no positive effects (Franke & Lieb, 2010, p. 858).

Glucose, our primary source of energy, serves as a comparatively effective memory enhancer. "Glucose memory facilitation effect" refers to the phenomenon whereby an increase in blood glucose facilitates memory function. Smith, et al. (2011) have reviewed studies on patients with memory deficits, in both old and young individuals, and demonstrated that the memory-improving effect of glucose is reliable, especially on verbal episodic memory. The underlying mechanism of enhancement is elucidated by the effect of glucose on cerebral insulin, synthesis of neuromodulator acetylcholine (ACh), potassium adenosine triphosphate (K_{ATP}) channel function, and brain extracellular glucose availability (M. A. Smith et al., 2011, pp. 779-781).

5.2 Brain Stimulation

5.2.1 Deep Brain Stimulation

Deep brain stimulation (DBS) is a common surgical therapy for treating neurologic and neuropsychiatric disorders (e.g., advanced Parkinson's disease, dystonia, depression, obsessive-compulsive disorder, and potentially Alzheimer's disease[27]). DBS modifies brain functions through the application of stimulation by implanting electrodes at specific sites in the complex neuronal circuitry underlying these functions; but the mechanisms resulting to the improvements in patients are not yet

[27] A recent study has shown that decline in glucose metabolism in temporal and parietal brain areas that is characteristic in Alzheimer's disease has been slowed down or partially reversed in early-stage patients after one year of brain stimulation of the fornix (G. S. Smith et al., 2012).

clear. Companies such as Medtronic Inc. and St. Jude Medical Inc. sell deep-brain stimulation devices (Cortez, 2012).

Despite its usage as medical treatment for managing the symptoms of disease, studies have shown that stimulating certain areas of the brain may lead to memory improvement. The ability to remember recently experienced facts and events requires interaction between the hippocampus and associated structures in the medial temporal lobe (e.g., entorhinal, perirhinal, and parahippocampal cortices). Studies in rodents have shown that electrical stimulations respectively of the perforant pathway (the neural structure projecting from the entorhinal cortex to the dentate gyrus and other fields of the hippocampus) and of the thalamus result in LTP, acetylcholine release, resetting of the theta phase, and neurogenesis in the hippocampus, which are all associated with memory improvement (Suthana et al., 2012, p. 503).

In human studies, those involving stimulation of the hippocampus have shown memory impairment or disruption (Suthana et al., 2012, p. 503). However, in one study, bilateral hypothalamic deep brain stimulation was performed to suppress the appetite of a patient with morbid obesity, and unexpectedly evoked detailed autobiographical memories of events that occurred more than thirty years ago (Hamani et al., 2008):

> Unexpectedly, the patient reported sudden sensations that he described as "déjà vu" with stimulation of the first contact test […] He reported the sudden perception of being in a park with friends, a familiar scene to him. He felt he was younger, around 20 years old. He recognized his epoch-appropriate girlfriend among the people. He did not see himself in the scene, but instead was an observer. The scene was in color; people were wearing identifiable clothes and were talking, but he could not decipher what they were saying. As the stimulation intensity was increased from 3.0 to 5.0 volts, he reported that the details in the scene became more vivid. (pp. 119-120)

A further study demonstrates that electrical stimulation of the hypothalamus modulates limbic activity and improves hippocampus-dependent memory function, e.g., recollection (Hamani et al., 2008, p. 122). Recently, a study reported that stimulation of the entorhinal region during learning improves spatial memory (Suthana et al., 2012). To be more precise, what this study shows is improvement in learning, which indirectly improves performance in behavioral tests.

There are concerns when it comes to using DBS as a means of memory enhancement. First, its effects on healthy individuals have yet to be examined. Second, as an invasive intervention with side effects (dependent on the stimulated location), it should only be used when the expected benefits outweigh the potential risks in serious conditions such as advanced Parkinson's disease.

5.2.2 Transcranial Stimulation

Transcranial magnetic stimulation (TMS) aims to affect neuronal activity through magnetic energy outside the head. The magnetic field generated by a coil induces electrical currents in the brain by means of activating synaptic inputs. TMS includes single or paired pulse TMS and repetitive TMS (rTMS). The former causes the site stimulated to depolarize and discharge an action potential. The effect only lasts for the duration of the stimulation. The effect of rTMS can last longer. It increases (high-frequency rTMS, simulation rates of >1 Hz) or decreases (low-frequency rTMS, stimulant rates of <1 Hz) the excitability of neurons depending on the intensity of stimulation, coil orientation, and frequency. It is currently used for diagnosis of neurological deficits and treatment of disorders such as depression and schizophrenia.

Unlike TMS, which induces focal currents in the brain, transcranial direct current stimulation (tDCS) induces brain polarization by delivering low current (1-2 mA) in brain areas of interest via two electrodes: stimulating (anode) and reference electrode (cathode). In anodal tDCS, an anode is attached to the scalp and a cathode to the contralateral supraorbital area, and in cathodal tDCS, the electrodes are reversely arranged (Brasil-Neto, 2012). Anodal tDCS leads to increased neuronal excitability, while cathodal tDCS results in decreased excitability (Floel & Cohen, 2007, pp. 251-252).

It is believed that TMS and tDCS can enhance or decrease plasticity in the cerebral cortex by inducing LTP- or LTD-like phenomena. The effects of such non-invasive brain stimulations are dependent on the brain area stimulated. One study has shown that false memories were reduced by 36% when low frequency rTMS was applied to anterior temporal lobes (ATL) after the encoding phase (Boggio et al., 2009). It is believed that when the left ATL is damaged, patients lose their semantic memory and the ability to name or label objects, while retaining the ability to retrieve literal details; thus it is hypothesized that disruption of the ATL will reduce false memories by diminishing our reliance on reconstruction through semantic processing, which tends to increase the rate of false memories. In addition,

story recall and verbal short-term memory tasks can be enhanced by rTMS (Floel & Cohen, 2007, p. 256). TDCS was reported to enhance memory consolidation when applied during slow wave sleep, and to improve recall of names of famous people when anterior temporal lobe tDCS is applied (cited in Dresler et al., 2013, p. 536).

Compared to DBS, TMS and tDCS are more feasible memory enhancers because they are not invasive and only few side effects were reported in healthy subjects: headache and local pain from TMS (Dresler et al., 2013, p. 536), while headache, nausea and insomnia were infrequently reported in tDCS (Poreisz, Boros, Antal, & Paulus, 2007). Moreover, portable TMS and tDCS are available. If their benefits on memory can be clearly established, they may be convenient devices for memory enhancement. It is suggested that non-invasive cortical stimulation, combined with training protocols, can enhance performance in certain tasks. Nevertheless, how they work on healthy subjects requires more studies to show their efficacy and safety.

5.3 Physical and Mental Exercises

5.3.1 Sleep

It has been reported that sleep is not only a physiological necessity, but also has positive effects on memory, especially the process of consolidation. Its influence varies with different kinds of memory: medium effects on declarative learning and significant effects on procedural or perceptual learning (Dresler et al., 2013, p. 533). It has been shown that after learning, the retention of declarative and procedural memory can be enhanced by sleep compared to an equal length of wake interval (Diekelmann & Born, 2010, p. 114).

The complementary functions of slow-wave sleep (SWS) and rapid eye movement (REM) are proposed to account for the mechanisms of memory consolidation in sleep (Diekelmann & Born, 2010, p. 123). The model assumes a temporary store and a long-term store. The latter can only learn at a slow rate but is able to sustain the information in the comparatively longer-term, whereas the former enables learning at a faster rate but only temporarily holds the information. (For declarative memory, the temporary and long-term stores are represented respectively by the hippocampus and the neocortex.)

After the new events are encoded in parallel in the two stores, the temporary store acts like an “internal trainer”: Its re-activation leads to concurrent re-activation in the long-term store, which enables the long-term store to incorporate the new

memories into its pre-existing network. To prevent interference, it is speculated that this "system consolidation" occurs during sleep (when no encoding takes place)—to be more specific, during SWS (Diekelmann & Born, 2010, p. 118). After the long-term store is trained, in the following stage of sleep, REM, it disentangles itself from the temporary store and undergoes "synaptic consolidation", which strengthens the memory representations that are re-organized in SWS. This is correlated at the synaptic level with pathways in the neocortex involving CaM kinases II and PKA, which are related to long-term potentiation (See §5.1).

5.3.2 Mnemonics

"Mnemonics" refers to "internal cognitive strategies aimed to enhance memory" (Dresler et al., 2013, p. 534). It is a process of associating to-be-remembered information with available systematic information, which could be internal (numerical/alphabetical order) or external (the setting of current environment). Three most common kinds of mnemonics, utilizing different forms of information, are methods for effectively remembering certain forms of information (Dresler et al., 2013, p. 534): "Method of loci" utilizes established memory of spatial routes, such as the way to work; "phonetic system", designed to help remember numbers, converts single digits to letters, then to words; "keyword method" aids learning a new language by associating words with similar pronunciations in languages already acquired. Compared to other forms of memory enhancement, such as pharmaceutical intervention or brain stimulation, mnemonics is prominently effective for extending memory retention and accuracy.

5.4 The Classification of Memory Interventions

There are different ways in which one can classify memory interventions. For instance, memory interventions can be differentiated according to the stage of the process at which they intervene. One can also classify them by which kind of memory they target (e.g., working memory, implicit memory, semantic memory, or episodic memory). Relevant to our consideration of the relation between the permissibility of memory enhancement and the property of the autobiographical self-model is a further method of classification: They can be sorted according to the way they are modified. This is shown in the following figure.

First, memory interventions can be categorized into *general* and *specific* memory manipulations: The former aims to manipulate general properties of the

memory system; that is, it influences all possible memories. Most memory interventions currently available belong to this category: Usually, their approach is to affect the whole system, for instance, herbs, nutrition, physical exercise, sleep and some pharmaceutical interventions (see §5). By contrast, the latter only targets memory with certain content and modifies properties of these memories. This is considered more difficult and complicated, but is expected to be available in the near future. For example, neuroscientists (Liu, 2012) have now succeeded in editing one simple content of mouse memory (see §5.1). Other methods, such as transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS), are also potential tools for intervening in one single memory (see §5.3.2). In addition, mnemonics, which has been developing for centuries, is also a promising technique (see §5.4.2).

Both specific and general memory manipulation can be further categorized according to which kind of property is modified: Accessibility, phenomenology, or quality. Let's first consider these as properties of general memory manipulations. To modify accessibility is to alter one's general ability to consolidate or retrieve the information: Theoretically, accessibility can be strengthened, weakened, or deleted. Excepting some pharmacological or herbal methods that controversially claim to be able to strengthen one's general memory capacity, there is no one acknowledged way to do this. Examples of such alteration are more easily seen in psychiatric conditions. Nevertheless, hyperthymesia, the condition in which subjects have superior capacity to recall autobiographical memory, can be seen as an example of strengthened memory. Weakened general accessibility of memory is commonly seen in aging subjects. Retrograde amnesia is an example of overall deletion. Phenomenology, on the other hand, refers to the phenomenal feeling that one has toward one's memory. Most of us have a feeling of familiarity, but such feeling is a matter of degree, and one can even lose that feeling (Klein and Nichols, 2012). "Quality of general memory" refers to the accuracy of the memory. As a general property of the memory system, accuracy refers to how well the structure of memory corresponds to real events in the past.

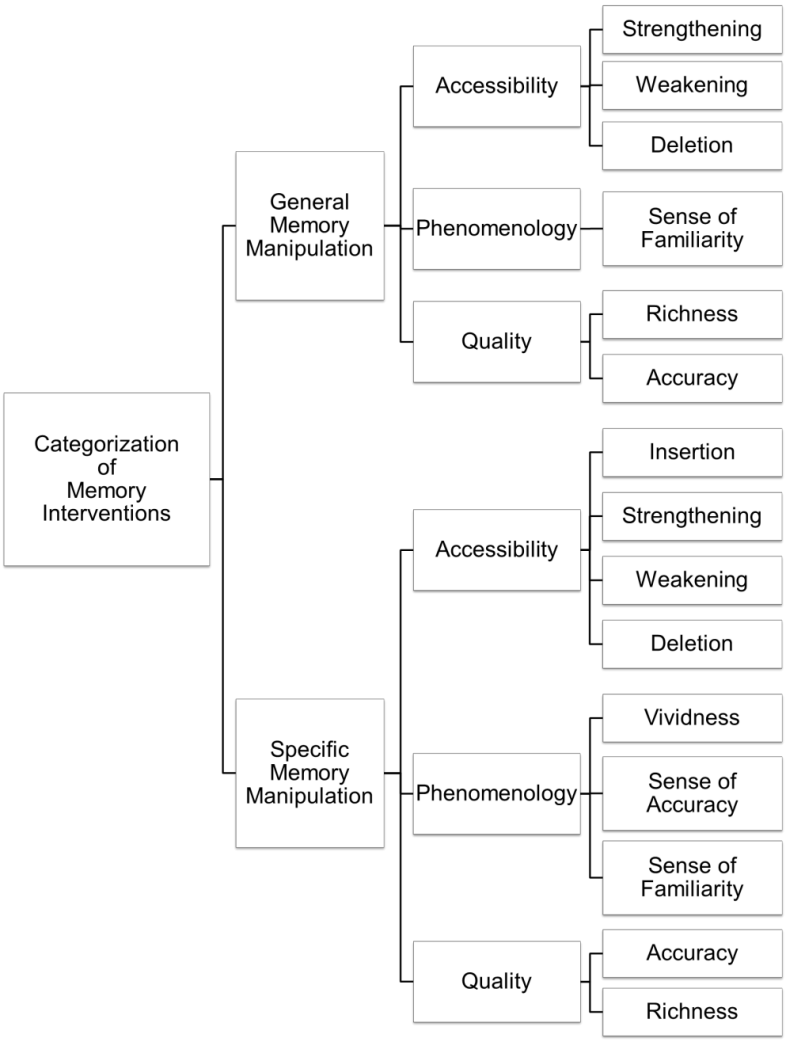

Figure 5. Categorization of memory interventions.

We can now consider specific memory manipulations. Altering the accessibility of a single particular memory can mean strengthening, weakening, deleting, or inserting that piece of memory. In fact, strengthening certain memories is what we intentionally or unintentionally do every day. Every time we recall something, the information is retrieved and then reconsolidated; reconsolidation strengthens the connections and increases its accessibility (see §2.2). Weakening and deletion are not conceptually difficult (since we forget most things), but is comparatively hard if one attempts this purposefully, for the best way to weaken a particular memory is to avoid reactivating its engram. But once one has a related

intention, the engram will be involved in relevant activation. This is a problem particularly for patients suffering from post-traumatic stress disorder (PTSD) or people haunted by certain unpleasant memories (e.g., the unhappy couple in *Eternal Sunshine of the Spotless Mind*); nevertheless complete deletion of certain memories is not yet available; at the moment the only available option is reduction of emotions tied to such memories. Again, it is easy to accidentally insert a memory. With some suggestion, our brain as a story-telling machine automatically fills the gaps in a story and creates a (vivid) false memory.

Second, the alteration of the phenomenology includes modifications of vividness, sense of accuracy, and sense of familiarity. The modification can make one feel that a memory is more or less vivid, accurate, or familiar upon recall. Note that phenomenology doesn't necessarily correspond to reality, and it can always just be a hallucination: When one feels that a memory is vivid or familiar, it doesn't mean that the information recalled is true, or that the event definitely happened. The feeling of accuracy does not necessarily correlate with the accuracy of the memory. Déjà vu is an example of the sense of familiarity without real recollection. Finally, quality of memory includes accuracy and richness of a memory.

As we will see in the next chapter, none of these memory interventions are directly linked to memory improvement or enhancement. What is seemingly considered a memory improvement (e.g., increasing memory accessibility or accuracy) does not necessarily contribute to one's well-being. For a memory intervention to be considered memory enhancement or improvement, it has to contribute somehow to a subject's self-interests.

5.5 Summary

In this chapter, different kinds of memory interventions are reviewed and their potential of being candidates of memory enhancers are assessed.

- *Studies on pharmaceuticals, genetic engineering, herbs and nutrition aiming at enhancing LTP have not shown any direct or consistent results in enhancing memory capability. Reasons for this are that (1) enhancing LTP does not necessarily improve memory and (2) more studies are required to prove its influence in healthy individuals (§5.1).*
- *Concerning brain stimulation, transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) may have potential as memory enhancers in the future. However, they are only beneficial*

when their application is combined with certain kinds of training. To predict or to demonstrate their effects, more about the specific roles each brain region plays in memory consolidation, retrieval, or reconsolidation needs to be known (§5.2).

- *Considered more effective are physical and mental activities such as sleep and mnemonics, which are thought to be less ethically problematic (§5.3).*

Then, for the convenience of discussions in later chapters, memory interventions can be classified by the ways in which memory is modified (see Figure 5).

- *They can be first categorized as general or local memory interventions: The former refers to the modification of the general property of memory, whereas the latter targets memory with certain contents.*
- *These two kinds of memory interventions can further be respectively sorted into the modification of accessibility, phenomenology, and quality: Accessibility refers to how easily information can be retrieved; phenomenology refers to the feeling that accompanies the recollection, such as the senses of familiarity and pastness; quality includes richness and accuracy.*

Chapter 6

A Fresh Look at the Concept of Memory Enhancement

6.0 Introduction
6.1 A Better Memory? The Conceptual Issues
6.2 Suffering and Memory Malfunction
6.2.1 The Function and Malfunction of Memory
6.2.2 Memory-Related Suffering
6.3 Self-Interests and Memory Enhancement
6.3.1 Self-Interests and Paternalism
6.3.2 Suffering and Autonomy
6.4 The Phenomenological Account of Memory Enhancement
6.5 Summary

6.0 Introduction

In §4, I delineated the concepts of improvement, treatment, and enhancement and proposed a phenomenological account. In §5, I introduced different kinds of memory interventions and a classification of memory interventions. The present chapter aims to apply the phenomenological account to the distinction between memory enhancement and treatment and to investigate the relevant conceptual issues in order to delineate the concepts of memory treatment and memory enhancement. The phenomenological account can also be applied to the inquiries of the treatment and enhancement of other cognitive functions.

According to the phenomenological account of enhancement (PAE),

> *Treatments* are, under standard circumstances, interventions that address a malfunction that results in an individual's suffering or potential suffering, and which the subject is not able to independently avoid or escape from.
> *Enhancements* are, under standard circumstances and without any unwilling suffering or potential suffering resulting from a target function, interventions that aim to manipulate the target function based on the subject's interests.

Nevertheless, this will require further conceptual clarification when we apply it to the distinction between memory treatment and enhancement. First, what do the normative concepts we attribute to memory really mean, for instance, expressions like "the memory is improved" or "a better memory"? This issue is considered in §6.1. The second section considers memory treatment: According to the phenomenological account, memory treatment is defined through the existence of suffering resulting from memory malfunction. Thus, the concept relies on a further concept of memory function and malfunction, and also on a relation between suffering and different kinds of memory malfunction. The subsequent section (§6.3) considers how memory improvement can be determined without reference to suffering. Related issues, such as the concern of paternalism and the relation between suffering and self-interest, are discussed. At the end of this chapter, I propose the *phenomenological account of memory enhancement* (PAME).

6.1 A Better Memory? The Conceptual Issues

No matter which view of the treatment-enhancement distinction one endorses, the concepts of treatment and enhancement (as well as "improvement") subsume the normative concepts of good and bad. There seems to be a direction, ranking, or criterion indicating a better or worse memory or cognitive system (see Figure 6). For instance, by definition, one's memory system must be (in some way) better *after* it is enhanced or improved. Therefore, when the treatment-enhancement distinction is applied to a certain "aspect" of a person, such as one's body condition, life span, cognitive faculties, skills, capacities, or morality, an account should be able to specify by which criteria it becomes "better" (or "worse").

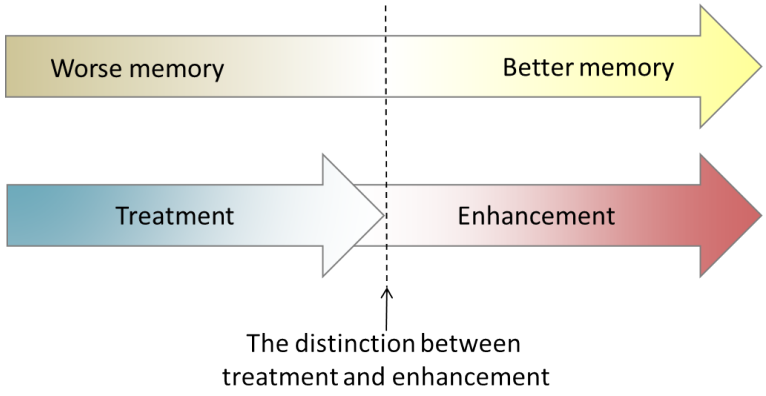

Figure 6. The model of memory treatment and enhancement.

The staff working papers of the President's Council on Bioethics ask, "what is a 'better' memory?" What is a good, or a better, memory, and what is a bad or a worse one? These questions are dubious for several reasons: (1) The term "memory" can refer to a variety of concepts; (2) it is not clear what the ethical concepts of "good" and "better" mean in this context; (3) it is unclear how a normative property can be attributed to memory.

The first issue was discussed in §2.1. The term "memory" can refer either to different kinds of memory (i.e., working, non-declarative, episodic, or semantic memory) or can be different understanding of memory (i.e., the neurocognitive capacity, the hypothetical store, the information stored, the retrieval of information, or the phenomenal awareness of remembering). Here, based on the classification of memory interventions in §5.5, I differentiate two notions of memory: The first refers to the neurocognitive capacity to encode and retrieve information, while the second refers to the content of memory. Note that these two aspects of memory are inter-related: The neurocognitive capacity can alter the content of memory and the content of memory affects one's capacity to encode or retrieve.

Next, how do we attribute normative properties such as goodness to a thing? When one says something is good, one could always be challenged with a question about why it is good. For instance, when one claims that reading is good, one may be asked why reading is good. One could answer that reading is good because it allows one to gain knowledge or information; nevertheless, this will be followed by a further question, namely, why is gaining knowledge or information good? Any answer will face the same question; to stop any further questions, one has, at one point, to postulate that a thing is good in itself. This leads to a distinction between intrinsic and extrinsic value. The concept of intrinsic value, a thing that is valuable "in itself" or "for its own sake", is to be distinguished from "extrinsic value" or "instrumental value", which is valuable "for the sake of something else" (Zimmerman, 2010). One may consider reading extrinsically good, for it enables one to achieve something intrinsically good. But the question follows: What is it that is intrinsically good?

Whether there is such a thing as intrinsic value, and what kind of things bears intrinsic value, is a meta-ethical issue that has been investigated by philosophers. G. E. Moore (1998) has offered a method for considering what might be good "for its own sake": "considering what value we should attach to it, if it existed in absolute isolation, stripped of all its usual accompaniments" (p. 91). For example, in adopting this method, a hedonist will conclude that pleasure is the only

good thing in absolute isolation. That is, according to hedonism, pleasure itself is the only thing that is intrinsically good or has positive intrinsic value, and suffering itself is the only thing with negative intrinsic value.

Then, what does it mean when normative properties such as good and bad are attributed to memory? Besides memory, terms such as "human enhancement" (Allhoff, Lin, & Steinberg, 2011; Bostrom & Roache, 2008; Bostrom & Savulescu, 2009), "enhancing human capacities" (Savulescu, ter Meulen, & Kahane, 2011) and "enhancement of executive function" are commonly used. What is a good human capacity, a good function, or better human? Take memory for example. One may say that one wants to have good memory, so that one can remember everything one has learnt.[28] Why is this super memory capacity considered good? Is it good for its own sake? Is it good for the sake of something else? What is it good for?

On the one hand, more is not necessarily good. Walter Glannon (2011) argues that strengthening one's long-term memory may result in more burdens than benefits (pp. 134–138). Increased retrieval of autobiographical memory will require more working memory to sustain it. If there is no proportional working memory, such strengthening will result in disturbance of other cognitive functions that also require working memory. Hyperthymestic syndrome is an example of this. Subjects undergo involuntary (as well as voluntary) recall of past experience. But this excess memory retrieval brings them difficulties, such as the problems of memorizing facts, of forgiving or getting over an emotional event, and of concentrating on current tasks. On the other hand, to have better memory doesn't necessarily result from having more memory. A study on amnesic patients with bilateral hippocampus damage shows that being able to utilize memory is determined not only by the accessibility of the information, but also by the capacity to retrieve information in a meaningful way (Hassabis et al., 2007). That is, what matters is not only *what* is retrieved but also *how* it is retrieved. As I will review in §6.2.1, the function of memory relies on its contribution to goals at higher levels.

What memory is good for is theory-dependent. As introduced in §4.1, this dissertation takes negative utilitarianism as its default practical ethical theory. In contrast to traditional (hedonistic) utilitarianism, according to which only pleasure is intrinsically good and only pain or suffering is intrinsically bad, negative utilitarians only contend the latter concerning the negative intrinsic value of suffering. If the only intrinsic value lies in suffering, anything else that is good or bad can only be

[28] Here, memory refers to the capacity to encode and retrieve.

such in an extrinsic sense: Something is good for the sake of reducing (overall) suffering, and something is bad for the sake of increasing (overall) suffering in the life of the subject in question.

According to the distinction between intrinsic and extrinsic value, and with negative utilitarianism, when we attempt to attribute a normative value, e.g., good or bad, to something, we are using it as an extrinsic value instead of an intrinsic one, for nothing except suffering can have intrinsic value: Other things can only be good in the sense that they are good for the sake of reducing suffering. Therefore, in this context, a cognitive function is good or better, if and only if it results in decreasing suffering. That is, an intervention *improves* a certain trait or cognitive function, if and only if after treated with the given intervention, such a trait or cognitive function is better in the sense that it reduces the suffering that resulted from the trait or function. That is, it is impossible to judge if something is good without consideration of its effect on the subject at the personal level.

So far, the reduction of suffering has provided an indication of what memory improvement means: If an altered memory results in less suffering, it is improved. However, two questions remain, which will be our focus in this chapter: (1) How is memory malfunction related to suffering? (2) How is memory improvement determined without suffering? The former and the latter respectively concern how memory treatment and enhancement are determined.

6.2 Suffering and Memory Malfunction

Cognitive treatment is distinguished from cognitive enhancement in that it aims to intervene with the malfunction of cognitive system, that has resulted in suffering on the part of an organism. That is, within the domain of memory treatment, a better memory is correlated with the reduction of suffering resulting from memory malfunction. In order to delineate the concept of memory treatment, we must look into the function and malfunction of memory and how memory malfunction can lead to suffering.

6.2.1 The Function and Malfunction of Memory

The function of memory was investigated in §2.3.2. This was based on the representationalism and the constructive view of memory, in contrast to the traditional view of memory function. According to the traditional view, which claims that the purpose of memory is to accurately represent facts and events

experienced, cases in which we are not able to correctly recall past representations are considered malfunctions. For instance, Schacter (2001) sees seven kinds of sins of memory as "memory's malfunction" (p. 4). Schacter's presentation of the pervasiveness of "malfunctions" in human beings suggests that these malfunctions may result from a false understanding of the function of the memory system, and he suggests that these misrepresentations do have adaptive advantages. Some follow-up studies have supported this idea and motivate a reconsideration of memory function.

All biological functions serve the organism for present and future gain, including memory (Westbury & Dennett, 2000). Compared to perception, whose function is to represent the current external world as veridically (and/or economically) as possible,[29] the function of memory is to allow the individual to have greater behavioral flexibility, rather than accurately representing the facts or events experienced (Suddendorf & Corballis, 2007). For instance, memory is often retrieved differently according to context or current environmental structure (Anderson & Schooler, 1991). Therefore, whether a memory system functions successfully should be determined by whether it serves the organism in reaching its current goal.

Furthermore, the idea that the function of memory is to facilitate behavioral flexibility is supported by studies which show that memory systems allow an organism to imagine future potential and past counterfactual scenarios (De Brigard et al., 2013; Schacter & Addis, 2007a, 2007b; Schacter et al., 2008). These capacities, aided by memory systems, allow the organism to mentally decouple from current external or internal stimuli, and consider more "creative" alternatives, thereby increasing the behavioral flexibility of the organism.

The contribution of memory to the achievement of the current goals is to be found in different ways. For instance, as what we traditionally consider, memory systems serve to retain information over time and to retrieve that information accurately later. There are situations that require us to remember things as correctly as possible, such as remembering a telephone number, or lawyers memorizing the law articles, technicians remembering the right measurements and procedures, and wine tasters remembering tastes of wine.

In addition, it can also be found in how memory affects other cognitive functions. For example, memory (i.e., episodic simulation) is shown to reduce

[29] Even in cases of perception, misrepresentation can be successful (Zehetleitner & Schönbrodt, 2013).

temporal discounting, the phenomenon whereby human beings discount the value of future rewards over time (Peters & Büchel, 2010). This not only prevents one from being limited to immediate rewards, but also allows one to make long-term plans. Boyer (2008) argues that the reduction of temporal discounting resulting from mental time travel enables individuals to cooperate, for most cooperation requires immediate contribution while resulting in long-term rewards. Accordingly, memory and other cognitive functions jointly contribute to the capacities of planning and cooperation. Furthermore, memory also plays an important role in our moral behavior. Its contribution to scenario construction affects our moral judgments. For instance, Peter Singer (1997) questions why we have the tendency to rescue a child who is drowning in front of us rather than helping a child who is starving to death somewhere far away. There is another example visible in the trolley problem: Compare the standard trolley problem with a version containing a fat man. In the first case, the trolley is driving in a direction that will result in five people being killed, and you can pull a trigger to change the direction of the trolley to another track where only one person will be killed. In the second case, you are standing on a footbridge above the track, watching the trolley moving forward and about to kill five people. The only alternative you have is to push down a fat man who is standing next to you onto the track to stop the trolley. Many of those who believe that they have the obligation to pull the switch to kill one person instead of five in the first case believe that they do not have the obligation to push the fat man down the footbridge in the second case (Singer, 2005, pp. 349–340; Thomson, 1976). From the perspective of utility, there is no difference between these two cases. What makes people have different moral intuitions about them?

This has interested many philosophers and psychologists: While some have tried to justify the difference in intuition (Thomson, 1976), others have suggested that emotion accounts for different reactions toward "personal" (pushing off the fat man) and "impersonal" (pulling the switch) violations (Greene, 2009; Singer, 2005). Episodic simulation allows the individual to simulate the scenario and therefore to receive the emotional feedback of these two cases, and the different simulated emotional states result in the different moral decisions. Among two constructed ASMs—one with oneself pulling the trigger and the other with oneself pushing down the fat man—the former is more compatible with our current ASM than the latter, and thus there is a difference in the inclination to take action in these two cases.

One crucial function of memory is to sustain an ASM. The ASM, integrated into a phenomenal self-model, allows one to own a collection of properties that differentiate oneself from others. These include one's personal history, prospective future, personality, values, preferences, and so on. Owning an ASM is an important aspect: As we will see later, sustaining an autobiography or a life story seems to be indispensable for being a person.

These lead to a re-thinking of the function of memory. De Brigard (2013) argues for an account of memory function according to which, in order to determine the function of memory, one has to determine the mechanisms of lower-level activity, and how this contributes to higher-level functioning. Thus, to determine the mechanistic function of memory requires an investigation into the way that its components contribute to the system and then how memory contributes to the functioning of the organism, helping it to reach its goals. Based on consideration of the contribution to the higher level, successful remembering and misremembering could both lead to well memory functioning. Therefore, the distinction between memory function and malfunction is not equivalent to the distinction between remembering and misremembering or veridical representation and misrepresentation; instead, it is goal-directed: It depends on whether the functioning of memory systems successfully contributes to the goals of the organism.

If the views discussed above are correct, the next step is to investigate how a level of analysis can be the best framework for studying the memory system and its relation to phenomena at the personal level. Most memory studies have focused on how the lower mechanisms contribute to memory system, whereas comparatively less research has focused on how the mechanistic function of memory contributes to the higher levels—that is, how the function of memory is conducive to achieving the goals of the individual.

6.2.2 Memory-Related Suffering

As the result of the last section suggests, misrepresentations do not imply memory malfunction, whereas veridical representations do not imply proper memory function; what determines memory function and malfunction is whether memory functioning successfully contributes to goals at the higher levels. If this is correct, the relationship between memory functioning and the goals of the organism is crucial. In order to understand how memory malfunction results in suffering at a personal level, this relationship is to be examined.

In §4.4.3 I argued that suffering plays an important role in the PAE: Not only does its reduction act as an indication of memory improvement in the domain of memory treatment, but its existence also works as a demarcation between treatment and enhancement. That is, improvements are, under standard circumstances, interventions that address malfunctions of memory that result in an individual's suffering or potential suffering, and which the subject has no ability to independently escape from. Accordingly, understanding the relationship between memory and suffering is crucial to drawing the boundary for memory treatment. What kinds of suffering result from altered memory capacity or memory states? How does memory malfunction lead to suffering? This section aims to explore the relation between memory malfunction and suffering.

As reviewed above, misremembering does not necessarily imply memory malfunction: There are cases in which misrepresentations of past events lead to successful behavior in the respect of fulfilling the needs for reaching the goals of the individual. Accordingly, in this section, I focus on the representations and misrepresentations that fail to contribute to goals at the higher levels and thus lead to suffering or deterioration of well-being.

We have differentiated different kinds of memory modification resulting from memory interventions in §5.4, and the memory alteration that results in suffering can be differentiated in the similar way. Memory-related suffering can be differentiated based on two conceptions of memory. The concept of memory can be understood as memory function or memory content: The former refers to the capacity to encode and retrieve different forms of information; the latter is the content of the mental or phenomenal (self-)simulation. Accordingly, memory-related suffering can be classified as *capacity-related* or *content-related*: The former refers to the suffering that results from excessive or insufficient neurocognitive capacity of encoding and retrieval; the latter refers to suffering that results from certain kinds of content of retrieved information.

Memory capacity-related suffering is commonly seen in patients with retrograde and anterograde amnesia, dementia, or hyperthymestic syndrome, whereas memory content-related suffering is seen in post-traumatic amnesia, post-traumatic stress disorder, and dissociative identity disorder. It is noteworthy that content-related memory alterations that result in suffering can also be understood as memory alteration resulting from either an excess or insufficient capacity for encoding or retrieving certain information. However, the capacity- and content-related distinction only serves as a convenient way to differentiate these two distinct

ways in which memory is related to suffering. Content-related suffering can be seen as capacity-related suffering (see §5.4). Second, capacity-related memory alterations are, commonly, general alterations in the sense that the affected memories are not selective in content; in contrast, content-related alterations are related to the special property of the content, which leads to either the selective memory loss or involuntary retrieval.

While we can accept that patients suffering from these kinds of memory alterations actually "suffer", very few studies have investigated what kind of suffering they undergo and how suffering emerges from different kinds of malfunctions of memory. We know that suffering can result from long-term inescapable distress, anxiety, etc.; there might not even be physical pain involved (see §4.1). Does an amnesic or dementia patient suffers the way a cancer patient suffers? Do amnesia and confabulation cause patients the same sort of suffering? Studies of suffering have focused on the relationship between physical pain (in a narrow sense[30]) and suffering, but more is required to explore how cognitive impairment results in suffering. How different kinds of memory alteration lead to suffering is not clear.

A review of the international literature on the lives of dementia patients from their own perspectives gives different views on how the disease affects patients' lives. From 50 papers, de Boer (2007) has classified themes that may account for the suffering of dementia patients: losses/changes, relationships, care and assessment, feelings, life satisfaction, denial/avoidance, minimization and/or normalization, continue living and fighting back, compensating and coming to terms with disease. Here I suggest four ways in which memory alteration can result in suffering: These include (1) difficulties in coping with everyday needs; (2) pressure from social interaction and expectations; (3) mood disturbance; and (4) difficulty in maintaining an ASM. First, while misremembering is prevalent, memory systems still serve the function of preserving information over time and recalling it in a useful way. Memory malfunction, resulting from an inability either to encode or to retrieve information of both declarative and non-declarative memory systems, prevents one from being able to cope with everyday needs. This is often linked to capacity-related suffering. Amnesia and dementia patients suffer from this difficulty. Severe amnesia patients, who have lost the ability to encode or/and retrieve episodic or/and semantic memory, suffer from difficulties in coping with

[30]"Physical pain" refers to one kind of bodily sensation mostly involves a damaging stimuli.

everyday life, while they may still be aware of it. In this context, memory treatments are interventions that help memory processing to extricate a person from such difficulties and thus alleviate the related suffering. For the time being, no medication is available that can effectively reverse neurodegeneration and induce neurogenesis; however, some kinds of occupational therapy, which allow patients to learn how to use new encoded information to replace un-retrievable old information, is considered to be memory treatment, because it enables the subject to escape the suffering that results from frustration or problems caused by memory incapability.

Second, as we live in a social and moral community constituted by other social beings, memory plays an important role in supporting our social agency. Our memory not only serves to preserve social information, such as faces and names of people we interact with, but also inform us of the norms that allows us to engage in social interaction in an appropriate way. In addition, appropriate social interaction is also enabled by our capacity to mentally simulate future or counterfactual situations. Memory malfunction can lead to failure to maintain appropriate interaction, which may result in social exclusion. What's more, maintaining an ASM is crucial to one's well-being, and the ASM is generated through a construction process not only based on memory elements, but also on the way it is constructed and the elements it utilizes, which are easily influenced by the conceptions of others towards us. Consequently, suffering can emerge not only from struggles for appropriate social interaction but also from the difference between one's self-conception and impressions and expectations from others.

Third, too much memory recall can trigger negative emotions and lead to suffering. Posttraumatic stress disorder (PTSD) is an example of the latter. What causes the suffering of PTSD patients is not the inability to retrieve information but the potential traumatic content of the simulational state, which is triggered frequently and involuntarily. Preventive interventions, such as beta-adrenergic antagonist propranolol, given before the traumatic incident, or medications such as selective serotonin reuptake inhibitors (SSRIs) taken after the incident, are interventions that improve one's potential memory states by modulating their emotional content and hence are considered memory treatments.

Last but not least, memory allows us to sustain an ASM, and memory malfunction results in difficulties in sustaining an ASM. In §3.2, I discussed the significance of an ASM. An ASM serves as a lens through which we attribute meaning to persons, objects, and events that we encounter: how they are related to our past, present, and future goals. The relationship between well-being and the

accuracy of memory has been investigated (Jetten, Haslam, Pugliese, Tonks, & Haslam, 2010; S. E. Taylor & Brown, 1988): What contributes to one's well-being is not the accuracy of memory but the internal coherence. As I have suggested, there are three characteristics of ASM—synchronic coherence, diachronic coherence, and global veridicality—where only synchronic coherence is directly linked to one's psychological well-being. Mental disorders have served as instances to show that our systems constantly maximize the synchronic coherence of our ASMs. There seems to be a kind of suffering resulted from the inability to form a synchronically coherent ASM. This can be found when the confabulated stories of Alzheimer's patients are corrected by external information, or when semantic dementia patients are unable to form the future dimension of the ASM while they are required to form future projections. Under such cases, memory treatments are interventions that help patients to construct ASMs.

To summarize, to determine if an intervention should be considered memory treatment, one has to investigate (1) if the intervention results in alleviation of suffering, and (2) if the suffering results from memory malfunction. As I have discussed, it is noteworthy that memory misrepresentation does not necessarily imply memory malfunction, while veridical memory representation does not necessarily imply successful memory functioning. The distinction between memory function and malfunction relies on its contribution to goals at a higher level. In addition, an intervention is not considered memory treatment without an appropriate context, including the condition of the subject and the relation between memory malfunction and suffering. What we consider memory treatment may differ from one case to another.

6.3 Self-Interest and Memory Enhancement

According to the PAME, memory enhancement refers to interventions that intervene with one's memory under condition where there is no memory-related suffering. The question then arises: What determines the improvement of memory? As I have argued above (§6.2), under the domain of memory treatment, when one experiences suffering resulting from memory malfunction, memory improvement is indicated by the decrease of memory-related suffering; however, what is memory improvement when there is no suffering? I propose that without suffering, the concept of memory improvement relies on one's self-interest. I will address some relevant points in this section, including the concern of paternalism and the importance of self-interest.

6.3.1 Self-Interest and Paternalism

As discussed in §4 and in the previous section (§6.3.1), suffering acts as an indicator for the distinction between treatment and enhancement, and for the determination of a better memory. However, what if there is no suffering, what marks memory improvement beyond treatment? Besides the elimination of suffering, what makes a memory better? How should we consider a memory enhanced?

From the evolutionary perspective, survival and reproduction are the most important individual goals that guide our choices and actions; that is, they might be the main indication of what is better for an individual. For instance, an animal equipped with the capacity to remember a similar situation in which it was in great danger will have better chance of survival and reproduction. This sort of trait is therefore retained in evolutionary history. Nevertheless, though survival and reproduction, as an indication of what is better for an organism, may explain most of animal behaviors, a variety of phenomena have shown that they are no longer the only goals that govern the lives of human beings.

One counterexample is homosexuality. Philip Kitcher (1997) has illustrated the issue as follows:

> Imagine that the recent suggestion that some male homosexuals have smaller hypothalamic nuclei than heterosexual males is confirmed by further research and, indeed, that we discover that nuclei of a particular size range always cause the men who have them to be exclusively homosexual. […] Nuclei of the smaller size appear to be dysfunctional, in the obvious sense that they would seem to cause their bearers to have fewer progeny. Yet even if this were so, would it be legitimate to intervene medically, to try to restore "normal functioning"? Are untreated men with the small nuclei "diseased"? (p. 213)

Other human behaviors have provided more counterexamples to Darwinian values, which refer to what promotes the organism's survival and reproduction. Take another example: More and more people or couples choose to be childfree or voluntary childless—fertile but with no intention to have children. The total fertility rate (TFR)[31] has decreased since the mid-twentieth century: from 2.53 in 1964 to 1.3

[31] TFR is the average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given fertility rate at each age (Central Intelligence Agency, 2013).

in 2006 in Germany, from 2.14 in 1973 to 1.26 in 2005 in Japan (as cited in Hara, 2008, p. 43), and from 7.04 in 1951 to 1.27 in 2012 (0.9 in 2000) in Taiwan (Department of Household Registration, 2012). Moreover, in 2003, 15.4% of German women desired no children (as cited in Hara, 2008, p. 52). Why did these women choose voluntary childlessness? Reasons against having children among childless women and men in Germany in 2003 include lack of a steady partner, wanting to maintain individual lifestyle, and concerns about not being able to enjoy current lives, the future of the child, the loss of leisure-time, possible conflict with job and professional activities, and the costs that the child may bring. Among these reasons, the precedence of leisure time, professional activities, and individual lifestyle, over having a child fails to conform to "Darwinian values".

The following extract, from Kitcher, illustrates such non-Darwinian values (1997):

> Even though we have been shaped by natural selection, we have non-Darwinian values, so that longevity and fecundity do not assume overriding significance for us. Human evolutionary history has equipped us both with the capacity for culture and with the ability to reflect on our predicaments and choices, and that combination loosens the connection between human valuation and natural selection, between what we want and what is good for our survival and/or reproduction. (p. 213)

Besides, more people perhaps choose to live for other ideals, such as art, human rights, and justice. We have seen countless cases of this in history: Take Hans and Sophie Scholl, who were active in the non-violent resistance movement in Nazi Germany, and Tibetans who self-immolate themselves. For these people, there are more important goals than their own survival or reproduction.

If "Darwinian values" are not the main goals of individuals, what might these goals be? Could there be an objective or empirical criterion such as a longer lifespan, a higher quality of life or more income, or normative criteria such as dignity, respect, or autonomy? Could it simply be a matter of taste? Is it even possible to find a universal criterion or a universal value that indicates the good that governs everyone's choices? These questions have interested axiologists for a long time. There exist two approaches to investigation: On the one hand, an empirical approach taken by moral psychologists investigates how human beings behave and reason, and aims to find out what drives a person to act in a certain way; on the

other hand, the moral approach asks what *ought* to be the value that dictates human action and how human beings *should* act.

The search for a single universal criterion of a better memory not only involves the issue of moral realism, but also concerns whether one should impose a value on others. Should one coerce others to do something or forbid them from doing something for their own good? In the context of memory enhancement, should we apply a single universal criterion of what a better memory is to everyone?

These questions lead to the issue of *paternalism*. Paternalism is defined by Gerald Dworkin (1972) as "the interference with a person's liberty of action justified by reasons referring exclusively to the welfare, good, happiness, needs, interests or values of the person being coerced" (p. 65). Nevertheless, Dworkin (1988) notes that when the concept is applied in any domain other than state coercion, it can be defined more broadly. Dworkin suggests defining paternalism as the violation of autonomy (p. 123), where autonomy refers to a capacity for identification with reasons for an action (p. 15). Such definition allows the inclusion of cases that involve no violation of liberty, such as a patient who is unwilling to be told about her condition, while the doctor insists on informing her of the truth for her own good (p. 122).

As such, the issue of paternalism concerns whether someone (X) other than the subject (Y) by doing or omitting something (Z) which interferes with the autonomy of Y for the reason of Z being good for Y. X can refer to government, states, parents, and so on. This issue often arises in the context of legislation concerning drugs, seatbelts, education, etc. (Dworkin, 2010). John Stuart Mill (2009) also regards paternalism as a danger to one's autonomy because it limits the freedom of an individual:

> [N]either one person, nor any number of persons, is warranted in saying to another human creature of ripe years, that he shall not do with his life for his own benefit what he chooses to do with it. He is the person most interested in his own well-being: the interest which any other person, except in cases of strong personal attachment, can have in it, is trifling, compared with that which he himself has [...]. (p. 129)

Mill argues that in most cases we are the best judges for ourselves. No one else can claim to know our interests better than we do. But is it true that we are always the best judges of ourselves? As we will see in the next section, we may be the best

judges of our current interests, but our capacities to judge for the future interests of ourselves may not be as good, or in certain cases, may be even worse.

As a result, concerning the issue of memory enhancement, what is a better memory and what kind of intervention is considered memory enhancement? There is no one universal value or goal for everyone, and one cannot define how memory functioning is better without considering the individual's goal and self-interests. Optimal memory functioning does not necessarily imply a better memory, so in some cases, being more forgetful might be better for the subject.

6.3.2 Suffering and Autonomy

Based on the lack of a universal value and the problem of paternalism, I propose that memory enhancement should be defined in accordance with the subject's values, preferences, and self-interests. However, one may argue that adopting suffering as the demarcation criterion for distinguishing eligibility for treatment may imply a paternalistic action. In reply, first, as I have addressed in §4.2.2, the distinction between treatment and enhancement functions as a "moral warning flag", which highlights the different kinds of moral issues that must be considered; it does not directly imply compulsion or permissibility. Second, memory treatment has an indirect link to medical necessity and consequently requires an objective measurable criterion, such as suffering. Third, there are exceptions to being one's own best judge of self-interest, and suffering is one of them: Suffering can temporarily or permanently change one's self-interests. Thus, in this section, I illustrate how and why self-interest under suffering should be treated differently.

Are we always the best judge of our self-interest? Can we be the best judges of our self-interest at a future time? The concept of *self-paternalism*, suggested by Rebecca S. Dresser (1981), concerns the issue of whether one is justified in making decisions about one's future for the future good, despite contradiction with future self-interests. In fact, we treat ourselves paternalistically all the time when we make decisions (e.g., the decision to take out a loan, the decision to go on a trip by car, and the decision to have another coffee). However, under most conditions, self-paternalism does not strike us as a problem, since we have a relatively diachronically consistent ASM: Our goals, preferences, values, and how we understand our past and future generally remain consistent. Consequently, it is natural that we identify with decisions made previously, because they serve the same goals with which we still currently identify. In addition, we have a sense of personal identity—the feeling that we are the same person as that at a former time—

which allows us to identify decisions made previously as *our own* decision. Even when we change our mind, we recognize that they were *our own* decisions, and therefore, we consider ourselves responsible for the decision and the consequences obtaining at later times.

However, the issue of self-paternalism becomes more evident in cases in which one's ASM is not diachronically coherent: That is, when the ASMs one constructs at two different times are incompatible. Such incompatibility can differ in degree: In cases in which two ASMs are less incompatible, one may have expressions such as "I don't know what I was thinking", whereas in cases which two ASMs are dramatically distinct, one may lose a sense of personal identity, and fail to recognize the other person as oneself. The latter can be seen in dementia patients or patients with dissociative identity disorder. The issue of self-paternalism arises as one's current interests contradict ones former interests.

According to Mill, we are only the best judges for our present interests, not our interests at the future time:

> [An] exception to the doctrine that individuals are the best judges of their own interest, is when an individual attempts to decide irrevocably now, what will be best for his interest at some future and distant time. The presumption in favour of individual judgment is only legitimate where the judgment is grounded on actual, and especially on present, personal experience; not where it is formed antecedently to experience, and not suffered to be reversed even after experience has condemned it. (Mill, 2004, p. 292)

One is the best judge for oneself, because self-interests emerge from one's ASM, which is directly informed by current perceptual input. However, as one's ASM can change over time, and can even change dramatically under certain circumstances, (e.g., extreme change of the environment), one's self-interests may change as one's ASM changes. To understand one's self-interests at a previous or later time, one has to simulate a counterfactual ASM in order to decide how another individual concerns one.

If neither classical paternalism nor self-paternalism is justified, this would seem to contradict negative utilitarianism: If a patient suffers from great pain but refuses to undergo treatment, which will (eventually) alleviate her suffering, according to negative utilitarianism, the treatment ought to be given; whereas to prevent paternalism, the patient's self-interest should be respected. What should be

done in this situation? Should the patient be treated paternalistically in order to ease their pain, or should the interest and autonomy of the patient be respected?

Dworkin (1972, 2010) considers whether paternalism is permissible, and under which circumstances paternalistic acts could be permissible. He questions whether a paternalistic act can do good to people against their will, because even though it may improve their situation in some respects, the process produces a form of the bad that outweighs the good of the treatment. Dworkin (1972) suggests that "we would be most likely to consent to paternalism in those instances in which it preserves and enhances for the individual his ability to rationally consider and carry out his own decisions" (Dworkin, 1972, p. 81).

I agree with Dworkin.[32] Concerning the tension between alleviating suffering and respecting the self-interest of the subject, I suggest that there is a difference in treating one's self-interest between situations in which one is undergoing suffering and in which one is not. As I discussed in §4.1.2, when one is undergoing suffering, one's working memory becomes limited because mental resources are occupied. Depending on the degree to which one suffers, one becomes less capable of (self-)simulation and is forced to focus on the current task of dealing with the pain or distress that results in the suffering. Imagine that you hurt yourself while you are working on a task. If you cannot get rid of the pain and suffering to a certain degree, you switch your focus from the task to thinking about how you can take care of the wound and alleviate the pain. Suffering has this property; it forces the subject to devote part of their mental resources to the goal of dealing with the cause of current suffering. Consequently, under such conditions, the ASM is constructed in a way to deal with the current need of alleviating the pain or what the suffering is pointing, and our limited cognitive and conscious access are used to deal with the information that is relevant to the suffering. Thus, interests resulted from this ASM are very likely to be distinct from those resulted from the ASM without suffering.

As such, I suggest that the practical solution to the tension between negative utilitarianism and the concern of paternalism is to treat cases differently depending on whether the suffering can be alleviated in a short period of time. If it can be

[32] Some (e.g., Conly, 2012) argue against autonomy and support paternalism based on the claim that Mill is wrong about people's competence in knowing one's self-interests (especially the means, how they get to their ends). However, such kind of argument has based on the assumption that a third person is capable of knowing what is best and the best way for the subject to reach the ends and ignores the diversity of interests.

alleviated, it is possible that one's current decision—made with the ASM with only a narrow workspace—can contradict with the future interests which results from a richer ASM. Thus, decisions for long-term plans should be made under the condition of no suffering. However, if there is no way to alleviate the suffering, one's current and future ASM are likely to be confined by the limited workspace. One's current self-interests should be immediately respected.

6.4 The Phenomenological Account of Memory Enhancement

Following the discussion above, the concepts of memory treatment and enhancement are defined as follows:

> *Memory treatments* are, under standard circumstance, interventions that address malfunctions of memory that result in an individual's suffering or potential suffering, and which the subject has no ability to independently escape from.
> *Memory enhancements* are, under standard circumstances and without any unwilling suffering or potential suffering resulting from the alteration of memory functions, interventions that aim to manipulate memory function based on the self-interests of the individual.

Different from the model I have illustrated in §6.1, this is shown below in Figure 7.

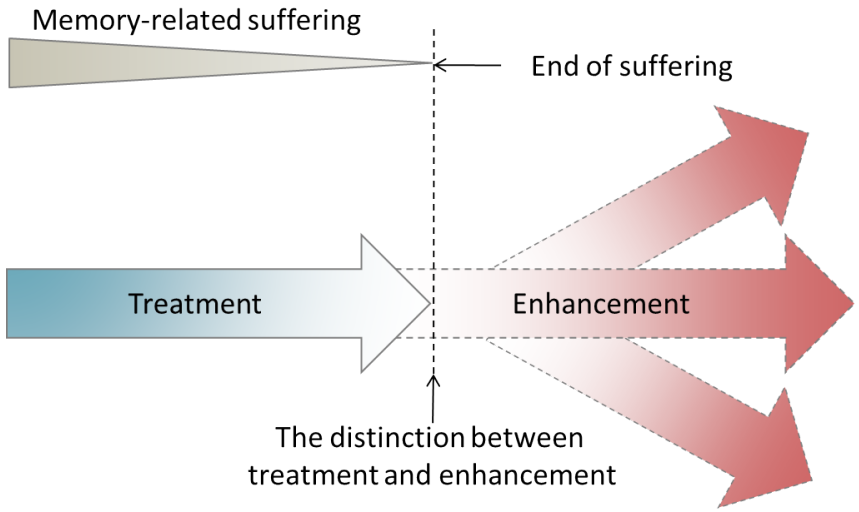

Figure 7. The phenomenological model of treatment and enhancement.

According to the PAME, the distinction between memory treatment and enhancement relies on the existence of memory-related suffering. An intervention is considered memory treatment if it alleviates the suffering resulting from memory malfunction, and an intervention is considered enhancement if, from the subject's perspective, it improves one's memory capacity and content. This is markedly different from the model sketched earlier. There is no one particular direction of memory enhancement that can be applied to everyone. What one person regards as enhancement is different from another: It depends on one's self-interest and how one conceives the intervention.

First, both the definitions of memory treatment and enhancement require an examination of the relation between memory functioning and the properties at a personal level, related either to suffering or individual goals. Second, what is regarded as memory enhancement or a memory enhancer relies on context, for instance, a pharmaceutical may be an enhancement for one person but not for another. Third, memory enhancement should not be restricted to pharmaceutical interventions. Evidence has shown that other kinds of interventions, such as brain stimulations or brain exercises, may be better candidates for memory enhancers (see §5).

Some may question: Is it possible that the improved memory is a hallucination? First, if one worries that we may have the phenomenology of having a better memory without a strengthened capacity of consolidating and retrieving information (e.g., we feel easier to recall certain information without realizing that the constructed memory is not veridical), conceptually it does not necessarily considered a hallucination. According to the phenomenological account, what an "improved", "enhanced", or "better" memory is depends on one's self-interests. An intervened memory system with increased general accessibility to past information does not imply that it is improved; it is not if it goes against the interests of the subjects. Second, if one wishes to have a false phenomenology, for instance the sense of vividness to the confabulated memory, it can be argued if such kind of modification if morally permissible, but it can be part of the subject's preference, and in certain cases (e.g., Alzheimer's disease), such hallucination help them stable their system. Third, since the phenomenology which is characteristic of memory can fail to correspond to the reality, we may be easier to mix up imagining and memory than we can detect by ourselves. What we should worry is that our preference may result from an ASM with low satisfaction of the functional constraint of global veridicality. It is worth noting that as we have discussed in §2.3, misremembering

can be beneficial to the subject. Nevertheless, we should bear the fact that we can be easily manipulated and have less autonomy than we would like.

6.5 Summary

To apply PAE to the distinction between memory treatment and enhancement, it is required to consider what determines a better memory.

- *The normative values of good or bad can only be attributed to memory when the conditions at the personal level are considered: either the existence of suffering or the self-interests of the subjects.*
- *With the existence of memory-related suffering, a better memory is determined by the reduction of suffering, whereas, without the existence of memory-related suffering, a better memory is determined by the subject's preference and identification of what a better condition is.*

Therefore, the phenomenological account of memory enhancement (PAME) is as follows:

- *According to PAME, memory treatments are, under standard circumstance, interventions that address malfunctions of memory that result in an individual's suffering or potential suffering, and which the subject has no ability to independently escape from; memory enhancements are, under standard circumstances and without any unwilling suffering or potential suffering resulting from the alteration of memory functions, interventions that aim to manipulate memory function based on the self-interests of the individual (see Figure 7).*

Because of the defining criterion of memory-related suffering in the distinction between memory treatment and enhancement, the concept of memory malfunction and its relation to suffering is crucial.

- *Memory function or malfunction is determined by the contribution of the memory processing to the goals at the higher levels (e.g., whether it allows the subject a greater behavioral flexibility).*
- *Memory malfunction can lead to suffering in a variety of ways: (1) difficulties in coping with everyday needs; (2) pressure from social*

interaction and expectations; (3) mood disturbance; and (4) difficulty in maintaining an ASM.

Chapter 7

Authenticity:
The Worry of the Loss of the Self and Identity

7.0 Introduction
7.1 The Concern of Authenticity in Cognitive Enhancement Debate
7.2 Authenticity as Self-Discovery
7.2.1 Authenticity as Self-Discovery
7.2.2 Criticism of Authenticity as Self-Discovery
7.2.3 Authenticity as Self-Discovery and the CE Debate
7.3 Authenticity as Self-Creation
7.3.1 Authenticity as Self-Creation
7.3.2 Authenticity as Self-Creation and the CE Debate
7.3.3 Criticism of Authenticity as Self-Creation
7.4 The Subjective and Objective Senses of Authenticity
7.5 Summary

7.0 Introduction

The idea of enhancing human traits through technology has stirred discussion of moral concerns in debate about cognitive enhancement (CE). One of the main concerns is whether CE endangers authenticity: This is the worry that CE results in the loss of self and identity. (For other ethical issues surrounding CE see §1, for a complete list of ethical issues of memory enhancement.)

In CE debate, two rival opinions are held by critics and proponents of CE: The former believes that CE alienates us, while the latter claims that it helps us pursue an authentic life. At first glance, it seems paradoxical that for both views authenticity is regarded as a moral ideal, a value that cannot be easily compromised and that should be taken into consideration when deciding whether CE is morally permissible or not. How can they, embracing the same ideal of authenticity, have different opinions on the influence of CE? This chapter aims to clarify different ways of understanding the concept of authenticity and to further elucidate them with the concepts of self, person, and identity reviewed and discussed in Chapter 3.

I will first describe the conflict between the two sides, which results from their different interpretations of the concept of authenticity (§7.1). Note that I mainly take Elliott and DeGrazia, respectively, as representative of the critics and the proponents of CE. Second, I take a closer look at their definitions and the background assumptions they endorse either implicitly or explicitly (§7.2-§7.3). In the next chapter, based on analysis of the concepts of self, person, and identity, and the idea of autobiographical self-model (ASM) in Chapter 3, I propose a framework in which we can understand the debate on authenticity through the properties of ASM.

7.1 The Concern of Authenticity in the Cognitive Enhancement Debate

In the debate surrounding CE, authenticity is one of the main issues often raised. The concept itself can be superficially understood as the moral ideal of being true to oneself:

> While the idea of authenticity has a complex history, the core of it is that we are authentic when we exhibit or are in possession of what is most our own: our own way of flourishing or being fulfilled. To be separated from what is most our own is to be in a state of alienation. (Parens, 2005, p. 35)

This is the value that is recognized by both critics and proponents of CE: An authentic life should be what everyone strives to achieve. This ideal aims to tell us what kind of life is worth living, and points to the way in which we should lead our lives. However, it is not clear what it means to live "in our own way". How do ways of living that are our own be distinguished from those that are not? The ambiguity of the concept leads to different interpretations. Although both critics and proponents of CE embrace the same moral idea of authenticity, each has a different opinion about how using CE technology influences one's life: The former believes that CE alienates one from one's "authentic self"; while on the contrary, the latter holds that CE can help us lead an authentic life. What underpins such a difference? As Eric Parens (2005) explains:

> [C]ritics and proponents […] of "enhancement technologies" share the moral ideal of authenticity, but they understand authenticity differently: they have different views about what it consists in, and thus about how to achieve it. (p. 35)

What results in the difference of the opinions of the two camps is the distinct conceptions of authenticity they endorse. This leads to different implications for how one should live one's life and what should be concerned: One considers the morally inviolable identity; the other worries about not changing one's life. The distinct conceptions of authentic life originate from different ideas of what one is.

On the one hand, critics of CE argue that enhancement technologies threaten authenticity or alienate us in the sense that they take away the subject's identity. For instance, the President's Council on Bioethics (2003) expressed their worry as follows:

> As the power to transform our native powers increases, both in magnitude and refinement, so does the possibility for "self-alienation"—for losing, confounding, or abandoning our identity. (p. 294)

Criticism of CE is based upon (1) the moral significance of being the same person over time and (2) the belief that CE technologies change traits or characters that are considered essential for being the same person. If one loses or alters a part of oneself, one does not remain the same person: One goes out of existence or, in other words, turns into another person. This is morally problematic; therefore, CE threatens authenticity.

On the other hand, proponents of CE argue that it enables one to achieve authenticity. Like the critics, to live an authentic life is to live a life that is one's own, but unlike the critics, proponents argue that a life of one's own is created rather than discovered. There exists no essence of one's self, and alternation the self is not necessarily morally problematic. Instead, what is morally unacceptable is to deny the possibility of "creating one's self", which is regarded as alienation. Based on the existentialist idea of authenticity, David DeGrazia (2005b) argues for a different idea: He holds a kind of essentialism based on animalism. We will look at this view in more detail in §7.3.

Excepting opinions of critics and proponents, the third conceptual possibility is that CE does not have any influence on authenticity. If this is the case, the concept of authenticity plays no role in CE debate. Authenticity can be neither a concern for the critics nor a reason for supporting the usage of CE. In the next chapter, based on my analysis of the different conceptions of authenticity, I will investigate what moral constraints we can posit for memory enhancement, based on the different accounts of selfhood and identity.

Moreover, whether and how CE affects authenticity differs from one kind of CE to another. Though considered CE, different kinds of interventions alter different aspects of cognition in different ways. Whether CE threatens authenticity or not depends on what kind of CE is in question.

7.2 Authenticity as Self-Discovery

The concept of authenticity concerns how we should live our lives and how we can come to live as full human beings. According to philosophers who conceive authenticity as self-discovery, with a view to achieving an authentic life one has to listen to one's inner voice or discover one's true self to know how to lead one's life. Carl Elliott is one such philosopher and has been addressing the topic of authenticity in CE debate. This section will focus on his ethics of authenticity: Starting with the origin of the idea in Charles Taylor, the following sections include a review of Elliott's ideas of authenticity and alienation, their relations with the concepts of self and identity, and their influence on CE debate. After a brief summary of his account and concerns, my criticism is presented.

7.2.1 Authenticity as Self-Discovery

Carl Elliott follows Charles Taylor's concept of authenticity. According to Taylor (1991), authenticity as a moral ideal refers to the notion that "each of us has an original way of being human" and a "measure" which determines what is our own "own way" (pp. 28-29). It originates from the individualistic idea in the Romantic Movement, which emerged at the end of eighteenth century when there was more freedom in society (p. 25). Before then, in Western pre-modern or hierarchical societies, where social roles were pre-determined, people worried about failing their lives or strove to live successful lives within the given frameworks. These frameworks are imposed by the external system and tell the members how they should live their lives in their specific roles. Only when people were freer in the society could each person start searching for a way of living that belonged to them personally.

However, in Twentieth Western societies, instead of worrying about failing to meet the demands of such frameworks, people encountered another issue—the issue of not knowing what framework underpinned their lives (Elliott, 1998, p. 179). Such uncertainty resulted in a feeling similar to vertigo for Taylor: "a sense of imbalance, because not only don't you know what kind of life to live; you don't

know what, if anything, can tell you" (as cited in Elliott, 1998, p. 179). This new predicament gave rise to awareness of the value of authenticity.

Back then, the original idea came from consideration of how one could tell right from wrong and good from bad, and being in touch with one's own feelings could be understood as "an instrument for moral knowledge and right action" (Elliott, 2003, p. 30). "Authenticity" came into existence from the idea that "the inner voice is important because it tells us what is the right thing to do" (C. Taylor, 1991, p. 26). It was the result of this shift, in the eighteenth century, of the origin of the sense of morality from the external (e.g., god or the Idea of the Good) to the internal—the voice within, which tells us what is the right thing to do. Instead of conforming to a moral guide and an external framework imposed upon us, authenticity emphasizes the importance of following one's "inner idea" or "true self". For Taylor (1991), the role of the inner voice as a moral guide then expanded to become the guide for how one should live her life in order to be a true and full human being. Taylor's notion of authenticity emphasizes the importance of not conforming to an externally imposed way of living; however, it fails to account for what it means to live authentically, as it is unclear what "inner voice" and "true self" mean for him.

Following Taylor's idea, Carl Elliott introduces his ethic of authenticity:

> The ethic of authenticity tells us that meaning is not to be found by looking outside ourselves, but by looking inward. The meaningful life is an authentic life, and authenticity can be discovered only through an inner journey. (2003, p. 35)

According to his ethic of authenticity (1998), first, life is a project, and it is a project of one's own in the sense that the meaning or significance of life depends on how one leads her life project, and that to a large extent she controls and is responsible for that project. Second, concerning the content of a good life, there is no one single and universal answer to the question of what a good life project is or how to achieve it: Because the content of a meaningful life project differs from one person to another, each person has to find her own answer. Furthermore, in order to search for an answer of one's own, one must look inward: Only through connecting with *one's true self* can we *discover* the way in which we can live a good and meaningful life.

The ethics of authenticity assumes that there is an internal framework, or what Elliott called "true self", that is to be "discovered". "Self-discovery", in contrast to "self-creation", which will be discussed in §7.3, is "an inner journey" (Elliott, 2003, p. 35). This idea relies on the existence of something inner that can be discovered. Nevertheless, what exactly is that something? Sex, personality, and

preference are examples in stories of becoming authentic that Elliott has illustrated (2003); that is, there are some properties essential to a person that should be discovered in order for them to be authentic. Before looking into his conception of authenticity, let's first consider alienation.

The concept of authenticity is opposed to the concept of alienation. According to Elliott (2000), a general idea of alienation is as follows:

> […] an incongruity between the self and external structures of meaning— a lack of fit between the way *you* are and the way you are expected to be, say, or a mismatch between the way you are living a life and the structures of meaning that tell you how to live a life. (p. 8)

It seems there might be at least two frameworks which can tell one how one is supposed to live a meaningful life: One is what is externally imposed on the individual by the community in which one lives, while the other originates internally from oneself. The meaning of one's life can be provided either by one or by both of them, and alienation describes the situation in which these two frameworks fail to match each other.

Alienation is further differentiated into three types by Elliott (2000): personal alienation, cultural alienation, and existential alienation. Personal alienation points to a mismatch between oneself and something from the outside (e.g., the social role one is expected to occupy). Cultural alienation, emphasizing the dynamic aspect, occurs when one is unable to keep up with changes in an external framework. But what is most conceptually interesting in the context of authenticity is existential alienation: It involves questioning the fundamental meaning of life, including one's form of life, one's values, and what makes a way of life meaningful (or pointless). When one starts to ask these questions and fails to find a framework that provides answers, this results in existential alienation, a radical form of disorientation. This is well elaborated in the following passage:

> […] not only don't you know what to do with your life, you do not know what could possibly tell you what to do. The structures that might have given life its sense and meaning are now contested or in question. The result is not just the feeling that you are ill-suited for your own particular form of life, or that your form of life is fading away; rather, it is a calling into question of the foundations of any form of life. Why this job, this church, this country, this house? Why this particular way of going on when I get up in the morning? Why *any* particular way? The result of these kinds of questions can be the sense that no form of life can really have the kind of justification that you feel you need. (Elliott, 2000, p. 10)

Existential alienation is a special kind of alienation:[33] It characterizes the loss of a framework. Not only is there a mismatch between the external and internal framework, but the existence of such a framework, whether internal or external, which plays the role of guiding a person to a meaningful life, is itself put into question.

What is the conceptual relation between "authenticity" and "alienation"? If, as was suggested earlier, an authentic life is a life that conforms to an internal framework, an inauthentic life might be a life that either conforms to an external framework, one that differs from the internal one, or conforms to no framework at all. In the first case, one may strive to match up with the external framework and thus live an inauthentic and alienated life. However, one may fail to be aware of one's internal framework: Through a variety of ways or for different reasons, one may be kept in the dark and never reflect upon or get in touch with oneself. (Self-deception may be an example of such case.) This could lead to an inauthentic life, but it is not clear if the person is alienated, for a "mismatch" requires an existent and contradicting internal framework. Whether one can be born and live without an internal framework is an interesting question.[34] In the second case, when one conforms to no framework, one is not only existentially alienated, but is inauthentic because not following any framework implies not following an internal framework. Whether one is alienated in the general sense or whether there is a mismatch between internal and external frameworks is irrelevant, because both of them are in a sense denied. Nevertheless, one may argue that existential questioning is part of an inner voice, or an internal framework. Can it be a framework, then, according to Elliott? This seems to create a paradox: To fundamentally deny the existence of any framework for a meaningful life could be a meaningful way of living life. However, this brings out the ambiguity of the idea of an internal framework.

Can internal frameworks be independent from external frameworks? What is the inner voice? What is a true self? In the following passage, I will examine two concepts that are often used by Elliott and are central to his idea of authenticity:

---

[33] It is not clear what the conceptual relation between the concept of alienation in general (the mismatch of the internal and external framework) and the concept of existential alienation. There can be a case where there is no mismatch between external and internal framework, which is characterized by alienation, but where there is existential alienation.

[34] It is probably easy to think of someone around you who blindly follows the mainstream and accept every arrangement or what she encounters without questioning. Is it possible for one to live in such way in every aspect? If it is, does it mean that one has no internal framework or that her internal framework is shaped to match the external one?

"self" and "identity". Unfortunately, this examination highlights the vagueness of his usage of the terms. In the next chapter, I will provide a better framework to understand what Elliott has tried to account for.

What do "self" and "identity" mean, for Carl Elliott? In his book, *Better than Well: American Medicine Meets the American Dream* (2003), Elliott outlines several stories in which people go on journeys in search of authentic lives and become true to their selves and identities. Jan Morris, a writer and foreign correspondent who underwent a sex-reassignment surgery to become a woman, commented on her transition, "[a]ll I wanted was liberation, or reconciliation—to live as myself, to clothe myself in a more proper body, and achieve Identity at last" (as cited in Elliott, 2003, p. 32). Elliott also writes, "Morris is not simply transformed; rather, her external male appearance is stripped away to reveal the true self that is underneath" (p. 31). The same idea is also present in the story of Sam Fussell, who identified himself as a bodybuilder and strove to become one, and Sandra, a Dutch woman who identified herself as a "small-breast type" and underwent a breast-reduction surgery.

First, the concept of the self plays a central role in Elliott's ethic of authenticity: To be authentic, one must connect to one's "true self" or "be true to oneself" (Elliott, 2003). What is *the true self*, and what is the contrary of such a self? When presenting positions shaping the debate on enhancement technologies, one tension Elliott brings up is between "self" and "self-presentation" (2003, p. 3). The former is what is felt from the inside, while the latter is what is presented to others (e.g., accent, status, and sexual characteristics). These two concepts are presented below as "the true self" and "the on-stage self":

> [T]he true self is the one that sits alone, a solitary self that endures over time, while the on-stage self is a mere persona, a type of useful role-playing that can be used or discarded as circumstances demand. (2003, p. 3)

There can be a gap between self and self-presentation, and it seems that this gap can be closed (p. 2). For instance, there is a gap for one who is born with male physical characteristics (self-presentation), but identifies himself as female (self), and through sex-change surgery, she can close the gap by changing her self-presentation to match her true self. Once her self-presentation matches her enduring self, her sexual characteristics become her "true" characteristics.

Furthermore, Elliott (2003) reminds us that the fact that people act differently under different circumstances or over time does not mean that they constantly transform into other people or have different selves. Instead, this simply shows that there are different aspects of the self (which I understand as "self-presentation"): Some tend to show themselves in certain circumstances while others in other situations (p. 49). He also notes that one may prefer one aspect of self (self-presentation) to another and thus let that aspect flourish.[35]

Elliott's distinction between "self" and "self-presentation" suggests that there is one part of self that lasts over time (perhaps throughout one's lifetime) and another that changes. Self-presentation is utilized to accommodate the demands of environment: Under different circumstances, different self-presentations are utilized. This is what others observe from a third-person perspective, and it explains why we often see people changing. However, the "self" can be different from self-presentation. If the two are matched, the character of self-presentation will become one's true character. Bringing in the distinction between internal and external frameworks, the internal framework is based on the self, while self-presentation can be regarded as the outcome of adaption to the external framework.

As for the term "identity", Elliott (1999) uses it as a constituent of how a person is and to a large degree, of a person's personality (p. 85). Though not explicit defined, his use of the term seems to respond to the question of who one is. Identity is something that determines who one is, and there is a set of characteristics that forms one's identity.

It is not clear if Elliott uses the concepts of true self and identity interchangeably. In his book (2003), where Morris' story of sex-reassignment is reported, he adopts Morris' use of terms such as "depriving of an identity" (p. 31) to illustrate how Morris thought about life before surgery, and "achieving identity" (p. 30) or "to achieve one's true identity" (p. 32) to describe the feeling Morris had afterwards. Prima facie, the concept seems to be identical to "true self". Nevertheless, Elliott also mentions that an identity is dependent on the recognition of others (p. 19). For him, the generation of identity cannot be solely inwardly generated and can only be developed through dialogue with others (p. 41). It is different from the idea of true self, which is considered inner and not influenced by outer factors. Thus, what Elliott calls identity can only emerge when the self-presentation matches the true self.

[35] What is interesting but is not explicitly discussed by Elliott is whether one prefers one aspect of self or one self-presentation to the others because it matches the true self.

7.2.2 Criticism of Authenticity as Self-Discovery

The ethic of authenticity that Elliott proposes is stimulating and points out the worry and concern that many people have towards enhancement technologies. Nevertheless, if one takes a careful look at his account, there is a lack of clear definitions of key concepts, such as "meaning (of living)", "framework", "true self", and "identity". To provide a clearer picture of his concerns, Elliott should explain to what exactly these concepts refer. The answers to these questions will give us a better idea of what authenticity as self-discovery means, and what critics of CE are worried about.

First, what is the true self, or in other words, what are the core properties fundamental to one? What is the criterion for one to remain the same person over time? According to Elliott's ethics of authenticity, which understands the concept of authenticity as self-discovery, to live authentically one has to discover and conform to an internal pre-given and static internal framework that is considered one's "true self". Therefore, the concepts of "true self" and "the self" play important roles in determining the concept of authenticity.

At first glance, Elliott's account of self and identity seem to assume "metaphysical essentialism", which holds that there is something necessary for one's existence (Metzinger & Hildt, 2011, p. 253). Nevertheless, he has defended the idea that "you can buy into the idea of an authentic self without buying into the idea of an essentialist self" (2003, p. 49). He provides an analogy with the concept of family: "An authentic self need not be defined by a single essential characteristic, in the same way that a family need not be defined by any single essential characteristic" (p. 49). Without further explanation, Elliott claims that "the ordinary, modern Western view of selfhood" is flexible enough to account for it.

To examine what the concept of "true self" could be, recall two kinds of concept of the self distinguished by Glannon (2007) in §3.1: Some use the term to refer to the concept involved in what results in the phenomenal self—the self-model and the functional constraints, whereas the richer concept includes the content of a phenomenal and an autobiographical self-model. If the first kind of concept of the self is adopted, the idea of authenticity as self-discovery would have to be built on the idea that ontologically there exists a "self" essential to an individual. This might be interpreted such that the self is "true" because it determines one's existence. However, there is no empirical evidence to support the idea of an ontologically existing thing that is a self.

On the other hand, Alexandre Erler (2011), who claims the existence of the "true self", explicitly considers the "true self" a form of narrative identity (Schechtman, 1996), that is, he uses the richer concept of the self (an ASM). However, if we acknowledge that narrative identity changes through time, we are required to provide criteria of which psychological characteristics are the core traits that would be problematic to alter. In addition, an argument is required for the claim that it would be morally problematic to alter these characteristics. Therefore, to understand the concept of authenticity as being true to one's self is conceptually problematic: We have to explain what the pre-given and static "self" refers to if we are not to endorse ontological realism of the self, and we must argue why it *shouldn't* be altered.

Second, what do framework and meaning mean? According to Elliott's ethic of authenticity, one's life is a project, and in order to achieve well-being, one requires a framework that reflects the meaning and value of life. Because every self is unique, there is no one universal framework. The only way for one to find one's own framework is through connecting to one's true self, that is, through self-discovery. Therefore, one leads an authentic life if one leads her life in a way that matches one's true self; otherwise, one is alienated from one's true self. Framework and meaning can be understood with the concept of an ASM. An ASM is a collection of mental (self-)simulations of past or potential future self-models of the system. ASMs allow us to know who we are and provide us with ways of knowing the relationships between oneself and objects or other agents. ASMs enable us to have preferences and values, and we are thus able to act and react according to our ASMs. That is, ASMs can be understood as "frameworks" which guide us through life, and meaning is what ASMs or frameworks provide us: It categorizes the world into desirable/undesirable, right/wrong, should/shouldn't, etc. The (self-)simulations let us comprehend things: how external (or internal) objects are relevant to us.

Third, Elliott's use of "identity" is conceptually confusing. On the one hand, he seems to use "identity" interchangeably with the "true self", which potentially refers to certain properties Elliott considers fundamental; on the other hand, as we will see later in the next section, one of his main worry about CE is built on the concern of losing one's transtemporal identity, which is a relation. He assumes that there exist core psychological properties that are necessary for one's existence. If these properties are altered, one may cease to exist or even transform into another person. Recall our discussion of transtemporal identity in §3.4; Elliott seems to

endorse the psychological approach of personal identity or worry about losing one's transtemporal personal identity.

7.2.3 Authenticity as Self-Discovery and the CE Debate

"Enhancement technologies" are defined by Elliott (2003) as being in contrast to therapy, in terms of for treatment. He refers to "the use of medical technologies not to cure or to control illness and disability, but to enhance human capacities and characteristics" (p. 27). However, it is not clear how he understands "illness", "disability", or "enhancement".

Why are critics of CE worrying about authenticity? According to Elliott (1998), there is a problem "only if it is not truly *your* life" (p. 182). Even if Prozac or other enhancement technologies give me a better personality, it is worrying if they alter my personality, because the result would not be *my* personality. For Elliott, what is worrying is not the likelihood of improvement but the fact that it may alter a person or change their capacities and characteristics, which are fundamental to the person's identity.

How CE endangers authenticity according to Elliott is ambiguous: It is not clear which the reason for the concern of using CE is, one's failure to match her true self or one's loss of her true self. When considering why using CE is problematic, Elliott (1998) answers that there is a problem "only if it is not truly *your* life" (p. 182). There are two ways to interpret this answer: It is problematic to alter my personality, (1) because it means my self-presentation will not match my true self, or (2) because it changes the personality which determines who I am and thus transform me into a new person. These two interpretations seem to be two distinct concerns, yet both raised by Elliott as concerns threatening authenticity.

The first case Elliott considers alienation: the true self remains intact, and the problem is that self-presentation does not match the true self. According to my understanding, interventions that merely change self-presentation should be less problematic. If an enhancement technology can alter self-presentation to match "the true self", it could free the subject from alienation resulted from the mismatch between "the true self". Could such kinds of CE exist? If there exist technologies that can modify one's core traits, a technology that modifies one's self-presentation—the characteristics that are not considered fundamental to one's existence by Elliott—should be more likely to exist and put to use. As this kind of CE help improve authenticity, it should be morally permissible, according to Elliott's ethic of authenticity. However, without further clarification of the concept

of "true self" or "core traits", it is difficult to come up with a criterion of what kinds of enhancement technologies are permissible and which are not.

In the second case, which seems to be more serious, personality is considered fundamental to oneself and is part of one's "true self", so altering that personality may lead to changing one's "true self". His worry is elaborated as follows:

> Much deeper questions seem to be at issue when we talk about changing a person's identity, the very core of what that person is. Making him smarter, giving him a different personality or even giving him a new face—these things cut much closer to the bone. And they cut close to the bone regardless of whether they are enhancements or cures, or even altering someone for the worse. They mean, in some sense, transforming him into a new person. (1999, pp. 28-29)

What is morally problematic for Elliott is the alteration of identity or changing oneself into a new person, and enhancement technologies or other technologies alike may put subjects under such risk.

(1) It is worrying if one's identity is altered.
(2) If the traits that are fundamental to one's existence are altered, one's identity is altered.
(3) Enhancement technologies may change these traits.
(C) Using enhancement technologies to change one's trait is worrying.

In this case, for the person in question, there is no such problem of losing authenticity, because the person goes out of existence and a new person comes into existence. There is no worry about self-discovery because the true self of the person in question is simply not there anymore. In the literature, the latter seems to be the main worry advocated by Elliott, but it has nothing to do with threatening authenticity: Based on Elliott's view of personal identity, it is senseless to say that one leads an inauthentic life if she does not exist anymore.

Apart from the worry of alienation, the other tension is between the natural and the artificial, or more broadly speaking, "what is given and what is created" (Elliott, 2003, p. 2). One part of what makes people uneasy about enhancement technology is that it is not natural. If one considers the "true self" as "given", it is perplexing that a person could identify with something artificial rather than

something natural. Nevertheless, it is worth questioning what it means by "natural" and how one can clearly distinguish natural from artificial. Besides, those who support the idea of authenticity as self-discovery and recognize the existence of a "given true self", they would have to explain what it refers to. The core psychological traits can be understood as a certain part of one's ASM; however, they have to argue why the part of ASM is fundamental and why other characteristics are not.

7.3 Authenticity as Self-Creation

Most proponents of CE comprehend the concept of authenticity as something closer to "self-creation", by contrast to "self-discovery". This section will focus on the idea of authenticity proposed by David DeGrazia, one of the main advocates of the idea of "self-creation" in CE debate, including the origin of the idea of self-creation, the examination of the concepts of the self and identity endorsed by DeGrazia, his concept of authenticity, and the implication in CE debate.

7.3.1 Authenticity as Self-Creation

Both Elliott and DeGrazia have taken notice of two traditions of authenticity, which are based on two different views of the self. The first, as introduced in §7.2, following the idea of the German Romantics, is the moral ideal that we ought to listen attentively to our "inner voice", which calls on us to live in a way that is distinctively ours. This idea presupposes that there is something called the true self that is "given" and is a being's essence. This "true self" cannot be changed, only discovered.

The other tradition, which is considered central to existentialism, holds a radically different view on the malleability of self. The existentialist concept of authenticity also suggests that we should be true to ourselves, and although this is superficially similar to the Romantic idea, the way of being true to one's self is here acutely different. Again, let us first consider the contrary of an authentic life. A life is inauthentic if one conforms to socially-approved roles and imposed values. From the existentialist perspective, by conforming to these or denying that one can do otherwise, one has given up and disowned oneself, and allows oneself to live under self-deception, which prevents one from facing the truth of what one really is (Guignon, 2004).

What then is the "real self"? Contrary to the Romantic idea of self or Elliott's distinction of self and self-presentation, existentialists not only deny that we have any pre-given essence or self, but also contend that there is nothing other than what we do in the world that defines us. According to Jean-Paul Sartre's view (as cited in DeGrazia, 2000, pp. 36-37), human beings are born without any "determinate nature" that we can discover. What we do and what we choose determines what we are. According to Existentialism, the norm of authenticity suggests that one should live her life as "the project of self-definition through freedom, choice, and commitment" (Crowell, 2010).

Whether an act is authentic or not is determined by whether I do the act "as myself" or "as anyone". That is, something I do is inauthentic if I do it because that is what anyone will do; on the other hand, if I do "is something I choose *as my own*, something to which, apart from its social sanction, I commit myself" (Crowell, 2010), I act authentically. This self-creative concept of authenticity is therefore closely tied with "freedom" and "autonomy": Do I have free will to choose and to act? Am I autonomous? For existentialist, the answer to these questions is not based on any external condition or factors but one's choice and identification.

Although advocating the concept of self-creation, DeGrazia endorses a different idea of self or different answer to the question of what one is to existentialists. According to DeGrazia (2005b), it is important that we have "essence" (p. 27), which is the necessary and sufficient condition for one thing to persist through time despite certain kinds of changes. The "essence" is the criterion of identity, which determines the condition under which one remains the same over time.

However, what is essential for DeGrazia? Unlike Elliott (see §7.2), he claims that it is not anything psychological: He has refuted "person essentialism", according to which we are essentially persons (see §3.2). According to person essentialism, any being that is once a person cannot be identical to a non-person at another time, and to be a person requires some psychological capacities (DeGrazia, 2005b, p. 30). Thus if person essentialism is right, I am neither the same person as the fetus which later contributed to my body nor the same being as the persistent vegetative state (PVS) patient that exists before the death of my body. Since the fetus and the PVS patient lack the psychological capacity necessary for being a person, and being a person is my essence, I cease to exist once these capacities are lost. Moreover, DeGrazia proposes another objection against person essentialism: It implies that we are not human animals. If a PVS patient is a human animal but not a

person, it is a case showing that an animal can be dissociated from being a person. Thus, being a person is distinct from being a human animal. It follows that if one is a person, then one is not a human animal.

Instead, DeGrazia adopts the biological approach according to which we are fundamentally "human animals, members of the species *Homo sapiens*" (2005b, p. 48). Compared to other traits that could be fundamental,[36] being *Homo sapiens* is the inviolable core characteristic. Yet one could question whether these core characteristics are threatened by the technologies that transform one into another species, such as gene manipulation. DeGrazia responds that the scope of the criteria can be larger than *Homo sapiens*, such as hominids, and the boundaries of our fundamental kind is what can be investigated and discussed (2005b, p. 278). Furthermore, the possible existence of such technology is debatable. The main point proposed by DeGrazia is that the essence and the criterion for numerical identity are biological, not psychological: We are essentially primates or members of some broader group of animals. Therefore, interventions that alter our psychological properties won't transform us into someone else as Elliott has worried about. In order to do so, they have to modify the core biological traits.

What are the concepts of authenticity and self-creation according to DeGrazia? How are they related with each other? According to DeGrazia, "[…] *any self-creation project that is autonomous and honest is ipso facto authentic*" (2005b, p. 112). There are three key elements of his idea of authenticity: "self-creation", "autonomy", and "honesty".

First, DeGrazia (2005b) uses the term "self-creation" to refer to "*the conscious, deliberate shaping of one's own personality, character, other significant traits* […]*, or life direction*" (pp. 89-90). Second, what does the concept of autonomy mean? For an action to be autonomous, one not only has to prefer the action one chooses, but one also has to identify the choice without considering it alienating:

> *A autonomously performs intentional action X if and only if (1) A does X because she prefers to do X, (2) A has this preference because she (at least dispositionally) identifies with and prefers to have it, and (3) this*

[36] Except the properties of being *Homo sapiens*, DeGrazia (2005b) also considers other candidates of core properties and refuted them one by one. These include internal psychological style, personality, and general intelligence, including memory, the need to sleep a certain amount of time, normal aging and gender.

> *identification has not resulted primarily from influences that A would, on careful reflection, consider alienating*. (DeGrazia, 2005b, p. 102)

That is, whether one's action is autonomous is only determined by the internal factor—one's identification and preference. Third, honesty is also important because even if one autonomously pursues a life, but systematically deceives others about who one really is, this self-creation project cannot be considered authentic for DeGrazia. The criteria of autonomy and honesty emphasize that one has to be true not only to one's narrative but also to others agents.

There are different accounts of the concept of autonomy. DeGrazia has endorsed coherentism, according to which autonomy is self-government in the sense that one governs one's action if and only if one is motivated to act because the motivation is coherent with the mental states relevant to that action (Buss, 2013). As a coherentist, Harry Frankfurt (1988) endorses a hierarchical account, according to which persons, by contrast to other animals who only own "first-order desires" (i.e., they want to or not to do something), are capable of generating "second-order desires or volition" (i.e., persons can want to or not to have certain desires or motives) (p. 12). One is autonomous with respect to an action if one's first-order desire for the action is sanctioned by a second-order volition that endorses the first-order desire (pp. 12-25). According to this account, one's identification with a desire for action makes one autonomous in undertaking the action.

7.3.2 Authenticity as Self-Creation and the CE Debate

Following Juengst (1998), DeGrazia defines "enhancement" as "interventions designed to improve human form or functioning". Treatment is defined by terms such as disease, impairment, illness, and departures from normal functioning, and is determined by reference to prevailing medical understanding. Enhancement, or interventions to improve human form or function that do not respond to "genuine medical needs", can be identified by the goal of improvement in the absence of medical need (DeGrazia, 2005b, pp. 205-206). Accordingly, "enhancement technology" refers to certain technology when employed for the purpose of enhancement.

As a proponent of CE, DeGrazia raises two problems for Elliott's ethic of authenticity (see §7.2) and their implications for using CE. First, as I have discussed earlier, Elliott's ethic is based on a false idea of the self. Self-discovery—being true to oneself and presenting oneself to others the way one is—which is the way Elliott

understands the concept of authenticity, presupposes that there exists some psychological traits that are "true" or "essential", and there to be discovered. Yet according to DeGrazia, there is no violable psychological trait and the only inviolable trait is being human animal.

Second, DeGrazia introduces the distinction of numerical identity and narrative identity made by Marya Schechtman (see §3.3), and further claims that Elliott's worry about losing identity results from equivocation of these concepts. The former concerns the reidentification question, which concerns the conditions one being at a time is the same as a being at the other time, while the latter concerns the characterization question, which asks which psychological characteristic, experience, or actions are properly attributable to a person. According to DeGrazia, what is morally problematic is altering numerical identity, not narrative identity. On the other hand, what is modifiable by enhancement technologies is narrative identity, not numerical identity.

DeGrazia then formulates Elliott's argument as follows:

1. Enhancement technology X alters a person's identity.
2. Altering a person's identity is highly problematic.

Therefore,

3. enhancement technology X is highly problematic. (DeGrazia, 2005b, p. 269)

Premise 1 is true if it refers to narrative identity; nevertheless, premise 2 is true if it refers to numerical identity. Therefore, we cannot conclude that enhancement technologies are problematic.

DeGrazia agrees with Elliott (see §7.2) that enhancement projects are connected to human identity, but the identity that he considers relevant is narrative identity, instead of the numerical identity with which Elliott is concerned. Consider the argument against CE formulated by DeGrazia. Examine the argument when the distinction of numerical and narrative identity is taken into account: Let's begin with numerical identity. For DeGrazia, it is indeed problematic to alter a person's numerical identity; however, based on his biological approach, according to which one is fundamentally a human animal, altering one's numerical identity implies transforming a human animal into a non-human animal, which is considered impossible for any enhancement technology. Therefore, enhancement technology is not considered problematic. Next, narrative identity. This is the sense of identity,

which concerns one's self-conception, self-told inner story, and self-evaluation, and is most relevant to enhancement. It is what is modifiable by altering one's human form and function. Then, the question becomes whether it is problematic to alter one's narrative identity, or not?

What is the problem in changing one's narrative identity? According to DeGrazia (2005b), there is no reason to consider changing a person's self-conception or one's mental autobiography problematic if she autonomously consents to the alteration (pp. 234-235). First, the reason why Elliott and others may find this problematic is that such alteration will change the traits that are fundamental to a person. Nevertheless, they never explicitly point out what exactly those core traits are, or what the criterion of these core traits is. Moreover, their idea is based on a static self that is independent of one's choices. To argue for this, one also needs to show why conforming to this "true self"—assuming we know what it refers to—is morally more important than one's autonomous choice. Second, the other worry is that to alter one's mental autobiography may result in dishonesty. However, the problem of dishonesty is, in its very nature, morally problematic, and there is no additional moral problem following this.

7.3.3 Criticism of Authenticity as Self-Creation

Based on how we distinguish between numerical and narrative identity, what seems to be morally problematic is altering numerical identity rather than narrative identity. Those who understand the concept of authenticity as self-discovery have a concern about CE based on the wrong assumption that there is something core and not changeable in narrative. As I have pointed out, they have to argue why certain part of an ASM is morally more important than the others. According to DeGrazia, autonomously modifying one's ASM is morally permissible, and the worry concerning CE is no more than the moral concern about deception, which is itself morally problematic.

With regard to numerical identity, one's essence is being a human animal, rather than being a person or other psychological traits. As such, the technologies that change one's personality, character, or traits are not morally problematic, and there is no such enhancement technology that leads to transformation of a human animal into non-human, which would result in the loss of one's numerical identity. Concerning narrative identity, as long as one's project of altering narrative is autonomous and honest, there is no problem in changing it, as substantial modification of one's personality is not considered an issue according to DeGrazia.

DeGrazia's criticisms and application of the distinction between numerical and narrative identity clearly point to problems in Elliott's view of self and ethic of authenticity. I agree with his criticism of Elliott's idea of the "true self", however, I do not agree with his endorsement of the biological approach (see §3.3). First of all, DeGrazia's account seems to be built on the presupposition that one is identical to the fetus that later contributes to the body and also to the patient in PVS. This presupposition is intuitively correct. However, such a claim should be the implication derived from a metaphysical relation of identity, which is investigated through conceptual analysis and scientific studies, rather than from an intuition, and thus it cannot be an objection to a metaphysical claim. In addition, being intuitively correct has nothing to do with being correct, for intuition can be wrong and what is intuitive can change from time to time.

Second, the criterion of numerical identity is dependent on the necessary and sufficient condition of a being. Instead of traditional formulations of "personal identity", DeGrazia, like Olson (2003a, 2003b), use "human identity". It is worth noting that these are in fact two different questions. Consider a statue made of a lump of clay: The identity of the statue and the identity of the clay are two distinct issues. Furthermore, it is confusing when DeGrazia raises the question of what we are: "[W]hat are we most fundamentally?" (DeGrazia, 2005b, p. 8) It is worth noting that there are different readings of this question depending on what "we" refers to: There is a difference between the sufficient and necessary criterion of being a conscious being, a person, a human person, or a human animal (see §3.3). It is not clear which question he is asking when investigating what we are or what I am and further the question of identity. If he means a conscious being, which is more interesting, in my opinion, for it requires a very basic form of self-consciousness, namely the capacity of being an experiencer (see §3.1) which is not possessed by a stone or a table.[37] Similarly, a fetus at a very early stage or a PVS patient may also lack it. On the other hand, if DeGrazia considers a human organism, there is an interesting question as to why the loss of personal or human identity is important for me. In other words, why would I care if a human animal ceased to exist? Once I lose the capacity of being conscious, why would the continuous existence of a human organism matters to me? This leads to doubting the presupposition of both DeGrazia and Elliott, that altering one's numerical identity is problematic. This question will be discussed in the next section.

[37] Without the consideration of panpsychism.

7.4 The Subjective and Objective Senses of Authenticity

"Now I feel like a machine, I've lost my passion. I don't recognize myself anymore" (Schüpbach et al., 2006, p. 1812). This was said by a patient with Parkinson's disease after having been treated with DBS. What such an expression indicates is a subjective feeling of not being oneself, which is what Kraemer (2011) calls "felt alienation" (pp. 3-4). Such a feeling can also be seen in some children with ADHD and who are on medication:

> I didn't feel like myself when I was taking the medication
> I just felt suckish all the time
> I was too quiet; it wasn't me
> My friends said I wasn't myself; I didn't laugh. (Singh, 2012, p. 362)

Not all children on ADHD medication feel alienated; some still feel that they are the same person.

According to Kraemer (2011), authenticity and alienation are the mental states that create this kind of phenomenology. They are opposites and respectively indicate the mental states that can be expressed as "I feel like myself" and "I am not myself" or "I am no longer myself" (p. 3). Kraemer claims that there is a normative dimension to these mental states. We experience authenticity as something we ought to strive for, while alienation is experienced as something we ought to avoid.

What Kraemer illustrates is the subjective *feeling of authenticity*. Such a feeling points to one major difference between the two different concepts of authenticity that Elliott and DeGrazia respectively endorse: Elliot's favored concept is objective whereas that favored by DeGrazia is subjective. According to Elliott (§8.2), one's being authentic depends on whether one conforms to the inner framework that is essential to oneself. Hence, whether one is authentic or not is determined by an objective criterion that can be assessed from a third-person point of view. It is, thus, possible that one could *feel* authentic but nevertheless be, in fact, inauthentic. For instance, a patient with Obsessive Compulsive Disorder felt very happy with the treatment of deep brain stimulation, even though the symptoms are not reduced (Schermer, 2013).

On the other hand, according to DeGrazia (§8.4.1), one's being authentic is determined by one's preferences and identifications. These rely on how the action or intervention in question is related to other relevant mental self-representations. That is, one tends to identify with and prefer an intervention if it is well and easily

integrated into one's ASM without creating incoherence. This is also what results in the feeling of authenticity illustrated above. This feeling of authenticity emerges when one experiences the coherence of one's ASM, whereas one experiences a feeling of alienation when the incoherence of one's ASM is phenomenally accessed.

7.5 Summary

In this chapter, I have examined the conflict between the critics and proponents of CE and respectively reviewed their different conceptions of authenticity and background assumptions concerning self and identity.

First, Elliott endorses the concept of authenticity as self-discovery, and a psychological approach of our fundamental nature.

- *According to Elliott's ethic of authenticity, life is a project of one's own in the sense that the meaning and significance of life depend on how one leads one's life project. However, in order to reach a good life project, one has to discover her "true self" to live authentically.*
- *Elliott endorses the idea of the self that is felt from the inside and is pre-given and constant across time. To Elliott, the importance of one's identity relies on the maintenance of one's unchangeable true self: Once it is altered, one no longer exists as the same person.*
- *Elliott's concept of the self seems to refer to core psychological traits that determine one's existence. Thus, what is worrying is not the likelihood of improvement but the fact that it may alter one's true self by changing one's core properties, which are fundamental to one's identity.*

As for DeGrazia, by contrast, hold the concept of authenticity as self-creation and animalism.

- *According to the idea of authenticity as self-creation, to live authentically, one must do and choose as myself instead of as anyone else.*
- *For DeGrazia, utilizing CE is not morally problematic, because what can be altered by CE is one's narrative identity, which refers to psychological characteristics, experiences, or actions attributed to a*

person, but not to one's essence as a human animal, a change which would influence one's identity.

- *To be authentic, according to DeGrazia, one has to be honest and autonomous, and an action is autonomous if it is preferred by the subject of the action, if her preference comes from her identification with the action, and if this identification is not the result of anything that she considers alienating, upon reflection.*

Chapter 8

Memory Intervention:
A Means or a Threat to Authenticity?

8.0 Introduction
8.1 The Framework of Authenticity
8.2 The Relationship between Authenticity and Memory Manipulation
8.3 The Moral Value of Authenticity
8.4 Authenticity and Memory Enhancement
8.5 Summary

8.0 Introduction

Although the issue of authenticity is central to the cognitive enhancement (CE) debate, it fails to provide a clear framework for discussion about what exactly we should be concerned with, and does not give a clear idea either of what we should be worried about or what enhancement will be able to help us achieve. These problems are due to the differences in understandings of the concept of authenticity in play and views of selfhood, personhood, and identity that are endorsed by various parties. This chapter aims to provide a framework for the debate by characterizing the concerns involved with reference to the autobiographical self-model (ASM). Such characterization allows us to consider how these concerns relate to different kinds of memory manipulation, and enables us to examine the normative value of such concerns by investigating possible states of the ASM.

First, I look at three concerns—truthfulness, identity, and autonomy—that are involved in two conceptions of authenticity. These concerns can respectively be illustrated by the constraints of functional adequacy for the ASM—global veridicality, diachronic coherence, and synchronic coherence (§8.1). Next, I consider the two-way relationship between the concerns involved in authenticity and memory intervention: The former constrains the permissibility of the latter, whereas the latter can modify the criteria for autonomy (§8.2). Then, with the characterization of these constraints in place, I examine the moral value of these concerns to see if authenticity is a value worth pursuing (§8.3). Last, I utilize the

framework to consider if memory enhancement is a threat or a means to authenticity (§8.4).

8.1 The Framework of Authenticity

As I have reviewed in the previous chapter (§7), critics and proponents of enhancement technologies embrace different notions of authenticity. The two conceptions they endorse involve multiple concerns. This section aims to provide a framework of authenticity: By examining Elliott's and DeGrazia's conceptions and summarize their debate into concerns of truthfulness, identity, and autonomy. They I show that these concerns can respectively be understood through the three functional constraints of ASM. That is, the debate of authenticity can be seen as the consideration of how ASM should be shaped and manipulated. The framework of authenticity provides an alternative way to examine the issue.

Carl Elliott's (1999; 2003; §7.2) concept of authenticity has its roots in Charles Taylor's ethics of authenticity (1991) and advocates that we should not live our lives in a way that merely conforms to the framework imposed upon us, for this leads to alienation. That is, we should listen to our inner voice or go through an inner journey to find our own framework for a way of living. As such, his concept of authenticity as self-discovery asserts that to live an authentic live one ought to "discover" the internal framework, which Elliott refers to as the *true self*.

According to Elliott (1999, 2003), the importance of authenticity lies in conforming to one's "true self"; that is, to be able to live an authentic life, one ought to consult one's internal framework to find a way of living that is one's own. As Elliott's notion of "the self" refers to a mental autobiography, the idea of the "true self" implies not just that one ought to act according to one's mental autobiography, but also that one's constructed mental autobiography should correspond to one's personal history and experience. When Elliott (1998) considers someone in a predicament, who has the option either to remain aware of the predicament or to take Prozac to relieve her from that awareness, the former is considered to be far better because she stays truthful to her situation. Accordingly, we see here the *concern of truthfulness*.

Elliott's next worry about the use of enhancement technology is a concern that enhancement technology may impinge on a person's identity. As I discussed in §3.4, the transtemporal identity relation is distinguished from a *sense of identity*: The latter refers to a feeling of being the same person at present as someone at

another time, while the former concerns the metaphysical relation of identity. Transtemporal identity can be further differentiated into *transtemporal human identity* and *transtemporal personal identity*. The former considers the criteria for which one human animal at one time is identical to another human animal at another time, while the latter considers the criteria for how one person at one time is identical to another person at another time. Because of the conceptual differences between "person" and "human animal", the conditions of persistence between the two are different. Transtemporal human identity yields a biological criterion according to which two human animals are numerically identical (§3.4.3). Transtemporal personal identity, on the other hand, is determined by the psychological criterion, and this criterion is dependent on how other agents conceive the ideas of personhood and personal identity (§3.4.4).

Loss of identity is a worry for both Elliott and David DeGrazia (2005b); however, because the only persistent criterion for DeGrazia is transtemporal human identity, which CE is unlikely to affect, he does not consider identity an issue in the context of CE. Elliott, by contrast, endorses transtemporal personal identity, to the extent that alteration of core psychological traits can lead to the loss of this identity relation. As such, the last concern involved in Elliott's worry about using CE is the *concern of (psychological) identity*. According to his objections to CE, what is worrying is not that it enhances a person, but that it may result in changing a person's persistence conditions by affecting their psychological characteristics. That is, there is a worrying possibility that through enhancement one may lose one's transtemporal personal identity, i.e., stop being the same person.

On the other hand, as a proponent of CE, DeGrazia (2000, 2005a, 2005b), understanding the concept of authenticity as an idea closer to self-creation, argues that it consists of two components: truthfulness and autonomy (§7.3.5). With regard to the concern of autonomy, according to DeGrazia, who endorses Harry Frankfurt's (1988) hierarchical account, an action is autonomous if it is preferred by the subject of the action, if her preference comes from her identification with the action, and if this identification is not the result of anything that she will consider alienating after reflection. Therefore, an autonomous action is in line with the interests of the subject. Such interests are based on one's identification. Consequently, if one identifies and prefers a memory intervention, having the intervention is autonomous. This is the *concern of autonomy*.

In short, three concerns are involved in the discussion of authenticity. Elliott's conception of authenticity includes the concern of truthfulness and identity,

whereas the concerns of truthfulness and autonomy are involved in DeGrazia's conception. Next, I consider these concerns by characterizing them through the satisfaction of functional constraints of the ASM.

The concept of an ASM, as developed in §3.2.4, is that of a collection of mental (self-)simulations of the past or potential future states of the system. It is sometimes referred to as "narrative" (Gallagher, 2000; Schechtman, 1996) or "mental autobiography" (Damasio, 1999, 2010). ASMs allow us to know *who we are*: They inform the system of its current and historical relation to the environment. An ASM not only provides self-related information, but when it is integrated into the active phenomenal self-model (PSM), it becomes phenomenally and cognitively available: It provides the system with a personalized context according to which a person comprehends, values, and acts. For a system to own an ASM, it requires not only a self-model but also the capacity for simulation (i.e., offline activation constraint), which allows it to mentally and partially depart from the current situation. Such a capacity further allows it to construct an inner mental time, and the system can thus be embedded in its personal history. The ASM shapes one's general character, or the property that we usually refer to as personality—considered from a third-person perspective.

An ASM is supported by both kinds of declarative memory—episodic and semantic memory (§3.2.1). According to the constructive episodic simulation hypothesis (Schacter & Addis, 2007a, 2007b), episodic memory contributes to the construction of a past or future event simulation by providing representations of one's past experiences: A simulated model of the future or the past is created (or recreated). Moreover, recent studies have suggested that despite episodic memory, semantic memory also plays a crucial role in constructing an autobiographical event: As representations of knowledge of oneself and of the world, it serves to structure episodic representations and to form a structured ASM.

An ASM is constructed by the interplay of episodic and semantic memory, and three functional constraints characterize how a self-model becomes an ASM: synchronic coherence, diachronic coherence, and global veridicality. Synchronic coherence refers to the consistency of the contents of (self-)simulata of an ASM constructed at a particular time. It is the most forceful constraint, and the failure to satisfy this constraint can be detected in the phenomenal experience, as it leads to one's suffering. Diachronic coherence refers to the consistency between the contents of ASMs constructed at different times. It often occurs when there is a dramatic change of environment. Global veridicality refers to how faithfully an ASM

simulates past experience and events, that is, the degree of correspondence between the structural and functional features. These constraints may be in conflict with each other, and when they are, synchronic coherence is the last constraint to be sacrificed.

These constraints can serve as the criteria to account for the concerns involved in the issue of authenticity—the concerns of truthfulness, identity, and autonomy. First, the concern of truthfulness involved in Elliott's and DeGrazia's ideas of authenticity relates to the awareness of external and internal states and whether one's conceptions about oneself and the external world are veridical. It can therefore be characterized by the constraint of global veridicality. A globally veridical ASM allows one to successfully gain access to past (self-)representations. If the constraint of global veridicality cannot be satisfied, one loses contact with the reality of one's personal status and history, e.g., what has happened in the past or what kind of person one is. This occurs in the cases of self-deception and confabulation.

Accordingly, if one endorses the concern of truthfulness and objects to cognitive or memory interventions or enhancements on the basis that one may lose touch with the truth, any intervention leads to the modification of an ASM which is characterized by the decrease of constraint satisfaction of global veridicality, is considered impermissible. That is, any intervention that leads the ASM to lose the property of global veridicality is considered a threat to authenticity. On the other hand, what if there is a memory intervention that allows increasing the satisfaction of the constraint of global veridicality? For instance, what if it prevented one from self-deceiving oneself? For both Elliott and DeGrazia, if it could lead to the satisfaction of the other two constraints they respectively endorse, it would be considered a means to authenticity.

As for the concern of identity, as identity refers to the concept of transtemporal personal identity, the criterion of the identity relation is the connectedness of one's psychological characteristics. It can be characterized by the functional constraint of diachronic coherence. Diachronic coherence allows one to be seen consistently from a third-person perspective across time, e.g., to have the same personality, preferences, interests, and personal history, whereas diachronic incoherence leads to shifts from one personality to another (e.g., dissociative identity disorder). However, diachronic incoherence often does not concern one's experience of having an autobiography, or of having past and future expectations

from the first-person perspective, which is related to the constraint of synchronic coherence.

For Elliott, any intervention that results in diachronic incoherence is considered a threat to authenticity. However, diachronic coherence is a matter of degree, and throughout our lives we experience diachronic incoherence to a small extent. In order for Elliott to form a better idea of authenticity, he has to specify the criteria for identity. In other words, he needs to state the extent to which diachronic incoherence leads to loss of one's transtemporal personal identity.

Last, the concern of autonomy based on one's identification and preference considers the synchronic coherence of the ASM and acceptance of the intervention. That is, if one's ASM is incompatible with the idea of using cognitive intervention, one does not identify with it or prefer it. Thus, as long as an intervention is compatible with the ASM, the intervention is autonomous. As I will discuss later (§8.2), memory intervention can modify the criteria for autonomy.

Following the characterization of the concerns in line with the functional constraints of an ASM, these constraints can be applied to the two conceptions of authenticity proposed by Elliott and DeGrazia. First, for Elliott, whose conception of authenticity involves the concerns of truthfulness and transtemporal personal identity, to be authentic, one ought to maintain the constraints of global veridicality and diachronic identity. Thus, if an intervention results in an ASM that fails to satisfy these two constraints, it threatens one's authenticity, and is considered impermissible for Elliott. On the other hand, according to DeGraiza, two concerns involved are truthfulness and synchronic coherence. Therefore, for an intervention to be considered authentic in DeGrazia's sense, it has to be synchronically compatible with the current ASM and its resulting ASM is globally veridical.

This framework, considering the concept of authenticity with the functional constraints, provides a clearer way to look into the concerns involved, and to judge if the conceptions advocated by Elliott and DeGrazia. With this framework, one is able to determine, with concrete criteria, in which way an intervention is considered worrisome or worthwhile by the bioethicists who defend the value of authenticity. However, as I will suggest later, the term "authenticity", due to the ambiguity it results in, should be abandoned. Instead, adopting this framework allows one to have a clear idea how one is affected by an intervention, and assess whether it truly requires concerns.

8.2 The Relationship between Authenticity and Memory Manipulation

After the establishment of an alternative framework for authenticity, this section examines the two-way relationship between the constraints of the ASM and memory manipulation. First, the constraints are used to provide a framework for authenticity. The issue of authenticity, understood as the concerns of truthfulness, identity, and autonomy is considered in terms of constraints on memory interventions. As we have seen in the last section, these concerns are characterized by the constraints of global veridicality, diachronic coherence, and synchronic coherence. We can examine if a memory modification violates any of these concerns, and if the modification is a means or a threat to authenticity, by considering whether it satisfies any of these constraints of the ASM.

Recall the classification of memory intervention introduced in §5.4: Memory alterations can be divided into *general* and *local alterations*. A local alteration involves the modification of a certain part of ASM, or one particular episode (e.g., a traumatic event). The local alteration involves the modification of *accessibility*, *phenomenology*, and *quality*. The alteration of accessibility makes the targeted episode more, less, or not at all available by strengthening, weakening, and deleting related memories. The insertion of a non-existent memory might also be an instance of local alteration of accessibility. Second, features of phenomenology including vividness, a sense of accuracy, and a sense of familiarity can also be modified. Finally, the quality of one episode—its accuracy and richness—can also be the target of manipulation.

Local alterations only locally affect the contents of ASM and preserve the general structure. The satisfaction of the functional constraints can be affected only when the targeted memory is deeply connected to other memories and its alteration leads to synchronic incoherence, which then results in automatic general alteration. In most of the cases, diachronic coherence and global veridicality are sacrificed to preserve synchronic coherence. This may occur in different ways: For instance, memory biases, in mild cases, and confabulation, in severe cases, result from the maximization of synchronic coherence (Beike & Landoll, 2000; Conway, 2005).

Regarding general memory alterations, they include alterations of general accessibility, phenomenology, and quality. To modify accessibility is to alter one's general ability to encode or retrieve: Conceptually, memory can be strengthened, weakened, or deleted. Phenomenology refers to the feeling of familiarity. Quality refers to the accuracy of the memory. General memory alterations refer to the

interventions that affect the overall properties of the ASM; thus, they are limited by the constraints of synchronic coherence, diachronic coherence, and global veridicality.

Therefore, to determine whether a memory intervention is permissible, it depends on the expected result of the intervention as well as one's notion of authenticity. The framework of authenticity provides a way to examine. According to the framework of authenticity introduced in the last section, if one based on one's account of authenticity, contends that a certain concern, e.g., truthfulness, is inviolable, any memory intervention that results in memory alterations that violate the corresponding constraint, global veridicality, is impermissible.

On the other hand, memory manipulation can alter the requirement for meeting the criterion of the concern of autonomy. As I examined earlier, the concern of autonomy involves the constraint of synchronic coherence; that is, one can examine the synchronic compatibility between an intervention and a subject's ASM in order to consider whether an intervention is permissible when considering the issue of autonomy. However, memory manipulation that modifies one's ASM can make an intervention that seemed permissible impermissible, or the other way round.

For instance, intervening in one's mind with cognition modification technology would not be considered an autonomous act by a bioconservatist. As he doesn't identify and prefer any sort of cognitive interventions, this action is not compatible with their ASM. That is, given the ASM of this bioconservatist, there are established criteria that make an action autonomous, and cognitive intervention fails to satisfy the criteria. Nevertheless, if we suppose that there exist neural technologies that allow us to modify and design the contents of memory, memory modification can be used to change the ASM of the bioconservatist into the one that is compatible with the use of enhancement technologies, or even the one that allows the subject to embrace the idea of cognitive enhancement. Such transformation to turn one from a bioconservatist into a transhumanist, changes what is compatible and incompatible with one's ASM, and thus, the action of intervening in one's mind which was not considered autonomous can become autonomous (, or the other way round).

8.3 The Moral Value of Authenticity

The framework for understanding conceptions of authenticity through the constraints of the ASM provides descriptive criteria for understanding the concerns involved in the debate surrounding authenticity. However, the framework not only allows us to examine the permissibility of memory interventions with these criteria, it also allows us to examine the normative value of the concerns involved and the conceptions of authenticity respectively endorsed by Elliott and DeGrazia. This section considers the moral values of the constraints, based on negative utilitarianism—which is endorsed as the default practical theory of this dissertation (see §4.1).

To begin with, the constraint of synchronic coherence is directly linked to suffering (see §6.2). This link is visible in techniques that aim to maximize synchronic coherence, as well as in mental illness. For instance, outweighing, justification, and closure are used to maintain synchronic coherence (Beike & Landoll, 2000; see §3.2.5). Another example is dissociative identity disorder: In order to deal with incompatible self-simulations, two or more ASMs are developed to keep the coherence of each ASM (Humphrey & Dennett, 1989). If synchronic incoherence is itself sufficient for suffering, based on negative utilitarianism—according to which we ought to minimize overall suffering—synchronic coherence is the constraint that should be satisfied. Therefore, according to the framework of authenticity, the concern of autonomy is understood as the synchronic coherence of ASM. Autonomy should be treated as a necessary criterion for assessing the permissibility of an intervention.

As for the constraints of diachronic coherence and global veridicality, failing to satisfy these constraints can lead to suffering in two ways: (1) It can affect one's personhood; (2) it can put the constraints into conflict. In §3.3.4, I argue for a normative concept of personhood. The use of "person" as a forensic term refers to the condition of being capable of being a member of the moral community. That is, personhood provides one with a moral status, which indicates how we *should* treat each other, and brings with it the rights and responsibilities one possesses as a member of the community. If one is a person, one is subject to certain rights and responsibilities. Unlike a PSM or an ASM, the criterion according to which one can become a member of the moral community requires the mutual agreement of the members of that community. That is, becoming a person requires the recognition of oneself as a person by other agents. Hence, because of the requirement of other

peoples' recognition, the concept of person differs from community to community (or culture to culture) and evolves with changing the conceptions of members in the community. To be a person is important, for it involves expectations of treatment and questions about how one *should* be treated.

Although the criteria for personhood are different in different communities, diachronic coherence and global veridicality of the ASM are considered critical criteria for being a person with certain rights and responsibilities. First, global veridicality allows one to possess shared representations and simulations with other moral agents. This is reflected in our expectation of truthfulness from others. Second, diachronic coherence is particularly important for personhood. Because those of us that possess a synchronically coherent ASM consider ourselves to exist continuously across time, we understand others in the same way. We understand each other as persons persisting continuously; that is, we consider each other as being the very same people we have been previously. The issue of personal identity, hence, becomes intriguing for us not only in how we experience ourselves, but also in how we understand each other, with all the moral implications that come along with that. For instance, the idea of responsibility develops based on the presupposition that persons exist across time: One has to be responsible for what one has done, because one is the same person.

Both biological and psychological criteria are involved in our folk psychological conception of the identity relation of persons. In our everyday lives, we utilize both biological and psychological criteria; nevertheless, the latter seem to play a more important role and are more decisive. We interact with other people and consider them the same person first of all because of their appearance. But soon after engaging in conversation, we are able to recognize a person through their behavior. A pattern of behavior reveals what one's ASM is like. ASMs thus underlie our personal identity. A diachronically synchronic ASM allows us to act consistently and maintain a consistent character across time. A diachronically incoherent ASM can therefore lead to loss of one's personhood.

The other way dissatisfaction of the constraints of diachronic coherence or global veridicality can result in one's suffering is in putting these constraints in conflict with the constraint of synchronic coherence. As we constantly strive to maximize the synchronic coherence of our ASMs, an ASM that fails to satisfy the constraints of diachronic coherence and global veridicality is likely to confront incompatible information that results in synchronic incoherence. For instance, dementia patients, who lose a stable ASM, struggle to maintain an ASM that is

synchronically coherent, but neither diachronically coherent nor globally veridical. Hence, they can suffer from denial of external information or other agents.

Therefore, as I have discussed, failing to satisfy the constraints of diachronic coherence and global veridicality can lead to suffering. However, as opposed to the constraint of synchronic coherence, neither constraint is sufficient for suffering alone—they only result in suffering indirectly. Furthermore, there are situations in which a subject is better off when these constraints are not satisfied. For instance, self-deception can result in a subject's well-being and confidence (i.e., self-enhancement) or even a better performance.

Therefore, understanding the concerns of autonomy, identity, and truthfulness through the constraints of synchronic coherence, diachronic coherence, and global veridicality allows us to see how they are related to suffering and to assess their normative value. As the criteria for the permissibility of cognitive or memory intervention, only a failure to be autonomous is itself sufficient for arguing against an intervention. On the other hand, violation of concerns of identity or truthfulness is not necessarily morally problematic. Then, if we return to the two conceptions of authenticity, the moral value of authenticity is to be reconsidered: Based on Elliott's conception of authenticity in which diachronic coherence and global veridicality are involved, the problem of the violation of this value is not sufficient for the impermissibility of an intervention. As for DeGrazia's conception, constituted by the constraints of synchronic coherence and global veridicality, the inauthentic intervention is necessarily morally problematic.

8.4 Authenticity and Memory Enhancement

This section will investigate the relationship between conceptions of authenticity and the phenomenological concept of memory enhancement. Is memory enhancement a means or a threat to authenticity? I will show, as Walter Glannon (2011, p. 142) suggests regarding CE, that memory enhancement does not necessarily make us inauthentic and can be consistent with authenticity.

In §6, I developed the *phenomenological account of memory enhancement and treatment*:

> *Memory treatments* are, under standard circumstances, interventions that address malfunctions of memory that result in an individual's suffering or potential suffering that she has no ability to independently escape from.

Memory treatments are, under standard circumstances and without any unwilling suffering or potential suffering resulting from the memory functions, interventions that aim to improve memory function based on individual interests.

Two components are critical in this account of enhancement and treatment: suffering and individual interests.

The phenomenological account adopts suffering as the demarcation of treatment and enhancement. This adoption is based on the phenomenological concept of health and a health-based account of the distinction between treatment and enhancement. The latter distinguishes treatment from enhancement by the problems to which it responds (Juengst, 1998; §4.1). According to the former, one is healthy if and only if one, under standard circumstances, has the ability to avoid or escape from suffering or potential suffering, and one is unhealthy if and only if one, under standard circumstances, is currently undergoing suffering and is not able to escape from the situation, or is prone to potential suffering (§4.4). This account, therefore, heavily relies on investigation into the relation of memory to suffering (§6.2).

Then, an enhancement must be an intervention that meets an individual's personal interests (§6.3). First, there is no general criterion that could determine what would count as a better mode of life for all human beings. Equipped with complex mental capacities, human beings have evolved beyond the governance of Darwinian values of survival and reproduction. Second, even if one is able to find or predict a good way of living for an individual, the question remains whether this should be imposed upon that individual. One might try to avoid paternalism by emphasizing the significance of autonomy and identification of the subject. What matters is how enhancement affects the ASM.

Concerning the debate of authenticity between critics and proponents of CE, is memory enhancement a means or a threat to an authentic life? As we have analyzed the concerns and constraints in the earlier sections, the permissibility of a memory enhancement does not rely on the intervention or on the enhancer, but on the ASM—namely on the alteration of the general properties of the ASM.

First, considering Elliott's conception of authenticity, a memory enhancement can be either a threat or a means to authenticity. This relies on how a memory enhancement alters one's ASM: If the modification results in the dissatisfaction of the constraints of diachronic coherence and global veridicality, it becomes a threat to authenticity according to Eliiott. By contrast, if it leads to the

satisfaction of one of the constraints, it is a means to authenticity. In addition, a memory intervention can be neither a threat nor a means if it doesn't affect the general properties of the ASM (e.g., the local alterations). Therefore, this conception of authenticity is not in itself sufficient for objecting to memory enhancement. As for the conception of authenticity endorsed by DeGrazia, this involves concerns of autonomy and truthfulness. However, based on the phenomenological account of memory enhancement, memory enhancement is defined according to one's self-interests; that is, if an intervention is not autonomous, it is not an enhancement. Thus, the only concern left for memory enhancement is the concern of truthfulness. Likewise, depending on the influence on the ASM, the intervention can be a threat or a means to an authentic life, or neither.

As a result, and as I have shown, neither Elliott's nor DeGrazia's conception of authenticity is sufficient for arguing for or against the use of memory enhancement. The permissibility of a memory enhancement depends on the alteration of the ASM that results from the intervention. The framework for understanding authenticity through the constraints of ASM provides a clearer criterion for consideration.

8.5 Summary

With the analysis of the conceptions of authenticity endorsed by Elliott and DeGrazia (see §7) and the conceptual of ASM and its three constraints—synchronic coherence, diachronic coherence, and global veridicality (see §3.2), this chapter has provided a framework for understanding "authenticity" as constraint satisfaction. It not only allows us to examine the permissibility of an intervention, but also allows us to examine the moral value of two ideas of authenticity.

- *The content of the debate of authenticity can be understood in terms of the constraints of the ASM: The concept of authenticity as self-discovery endorsed by Elliott concerns the constraints of diachronic coherence and global veridicality, whereas the concept of authenticity as self-creation endorsed by DeGrazia concerns the constraints of synchronic coherence and global veridicality (see §8.1).*
- *The constraints and memory interventions exist a two-way relationship: (1) The constraints can be used to confine the permissibility of an intervention including memory interventions; (2) memory intervention*

by modifying the ASM can alter the criterion for autonomy (see §8.2).

- *The constraint of synchronic coherence of an ASM should be satisfied, because synchronic incoherence is sufficient for suffering. Neither the constraint of diachronic coherence nor the constraint of global veridicality is sufficient for the impermissibility of memory interventions (see §8.3).*
- *Autonomy is necessary; however, truthfulness or identity is not sufficient for confining the permissibility of memory intervention or memory treatment; that is, neither concept of authenticity is sufficient for arguing for or against memory enhancement (see §8.3).*
- *No matter which conception of authenticity is endorsed—authenticity as self-discovery or authenticity as self-creation—memory can be a means or a threat to authenticity, or neither. It depends on examination of alteration of the ASM. The framework introduced in this chapter allows for a clear examination.*

Chapter 9

Conclusion

9.0 Introduction
9.1 Summary of the Discussions
9.1.1 Memory
9.1.2 Memory Interventions
9.1.3 Selfhood, the Autobiographical Self-Model, and Personal Identity
9.1.4 The Concepts of Authenticity
9.2 The Phenomenological Account of Memory Enhancement
9.2.1 The Phenomenological Concept of Health
9.2.2 The Phenomenological Concept of Enhancement
9.2.3 The Phenomenological Concept of Memory Enhancement
9.3 Memory Enhancement and the Issue of Authenticity
9.3.1 The Conceptual Issue of the "True Self"
9.3.2 Authenticity as Constraint Satisfaction
9.3.3 The Moral Value of the Constraints and Authenticity
9.3.4 Memory Enhancement and the Constraints
9.4 Open Questions for Future Studies

9.0 Introduction

This dissertation aims to examine the normative issues of memory enhancement, and focuses on two issues: (1) the distinction between memory treatment and enhancement; and (2) how the issue of authenticity concerns memory interventions, including memory treatments and enhancements. The first issue questions how memory enhancement is distinguished from memory treatment. This question involves the concepts of health, illness, and disease. Based on a phenomenological account of health and illness, I argue for a phenomenological concept of memory enhancement. Second, the issue of authenticity asks whether memory interventions or enhancements endanger "our true selves". By analyzing different conceptions of authenticity and relevant concepts of self and identity, I propose that the ambiguous concept of authenticity should be understood in terms of properties of the

autobiographical self-model (ASM): synchronic and diachronic coherence and global veridicality.

This chapter is comprised of three parts: The first (§9.1) summarizes what I believe to be the chief results of the discussions presented in this dissertation. These are the theoretical foundations of my discussion of the two issues mentioned above. The second part (§9.2–§9.3) summarizes arguments for the phenomenological account of memory enhancement, and the framework of the ASM, to account for the issue of authenticity. The last part (§9.4) considers open questions for future conceptual and empirical studies.

9.1 Summary of the Discussions

9.1.1 Memory

The concept of memory is used to refer to different ideas of memory or different memory systems. In this dissertation, I use the term to refer to the retrieval of information (§2.1.1) and adopt a psychological taxonomy of memory systems with a distinction between non-declarative and declarative memory, where the latter can be further differentiated into semantic and episodic memory (L.R. Squire, 2004, p. 173; see §2.1.2). Episodic memory and semantic memory are distinguished from each other by characteristics of memory, including being (1) experience- or belief-like, (2) dependent or independent from context and self- or world-related, and (3) accompanied by a sense of pastness and familiarity or a sense of knowing. Nevertheless, what is more important—and has been overlooked in memory studies—is how these two memory systems interact with each other to create our conscious experience during recollection. Most of the recollections we undergo every day, including autobiographical memory (the recollection of one's self-related information) are realized by the interplay of episodic and semantic memory systems: The former provides episodic details; the latter acts as a schema to integrate them (Irish & Piguet, 2013).

This dissertation adopts a representational theory of memory (§2.2.2): Memory is comprised of (self-)simulata. In contrast to (self-)representations, which rely on current inputs and whose contents are stimulus-related, (self-)simulations triggered by external or internal stimuli do not involve the kind of activation of sensory correlates involved in mental representation, and the contents are not stimulus-correlated (§2.2.3). Two functional constraints are involved in the generation of a (self-)simulation: offline activation and phenomenal transparency.

The former allows for generation of (self-)simulata, which are independent of current internal or external states. The organism, by satisfying the constraint of offline activation, is capable of generating states such as recollection, planning, mind-wandering, and dreaming. Phenomenal transparency is the phenomenal property of a conscious state, which appears when earlier processing stages of the representation are not available for attentional introspection. But this is not an all-or-nothing property: Some states are more transparent; others are more opaque. A phenomenal (self-)simulatum tends to be more opaque than a (self-)representatum.

Memory studies have shown that the mechanism of memory is a process of deconstruction and construction (Schacter et al., 1998), rather than a Lockean kind of storage (Locke, 2008). This kind of mechanism is responsible for the memory distortion that is common in human memory. Studies on the constructive nature of memory have provoked us to rethink the function of memory. Against the traditional idea that the main function of memory is to faithfully represent past episodes or information, recent empirical evidence showing that memory contributes to future and counterfactual thinking (De Brigard et al., 2013; Schacter & Addis, 2007a, 2007b) provides support for the idea that the main function of memory is to allow the system greater behavior flexibility (Schacter et al., 2011; Suddendorf & Corballis, 2007; §2.3.2).

9.1.2 Memory Interventions

Memory interventions can be classified by the ways in which memory is modified. They can be first categorized as general or local memory interventions: The former refers to the modification of the general property of memory, whereas the latter targets memory with certain contents. These two kinds of memory interventions can further be respectively sorted into the modification of accessibility, phenomenology, and quality: Accessibility refers to how easily information can be retrieved; phenomenology refers to the feeling that accompanies the recollection, such as the senses of familiarity and pastness; quality includes richness and accuracy (see Figure 5).

Different kinds of memory intervention are overviewed in §5. They include pharmaceutical interventions (e.g., memantine and amapkine), genetic engineering, herbs (e.g., *Ginkgo biloba*), nutrition (e.g., glucose), brain stimulations (e.g., deep brain stimulation (DBS), transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS)), and mental and physical exercises. The effectiveness of these interventions is largely dependent on the individual baseline

performance; that is, current memory interventions are more effective as memory treatment than as memory enhancements. What seem more plausible are brain stimulations and mental exercises such as mnemonics. Among brain stimulations, TMS and tDCS are more feasible as memory enhancers because they are not only non-invasive and show few side effects in healthy subjects, but are also available in portable versions. Comparatively, they are less expensive and more easily integrated into one's everyday life.

9.1.3 Selfhood, the Autobiographical Self-Model, and Personal Identity

We have the phenomenal experience of being someone who is the "center" of *our* world and the recipient of all sensation. This phenomenal experience of being an experiencer—the phenomenal self—results in an intuition of the realism of the self. However, the phenomenology of substantiality does not suggest any metaphysical realism of the self (Metzinger, 2011) and no empirical evidence can support the ontological existence of such a thing as a self.

Metzinger (2004) proposes a naturalized theory—the Self-Model Theory of Subjectivity (SMT)—to account for phenomenal properties with representational and functional properties, and shows how a phenomenal self can emerge from the information-processing system by satisfying a series of constraints: If a self-model—a coherent internal model of the system as a whole—satisfies the constraints of *globality*, *presentationality*, and *transparency* (see §3.1.4), it becomes a phenomenal self-model, which allows its contents to be consciously and cognitively accessible and results in the emergence of the phenomenal self. However, phenomenal subjectivity only arises after *phenomenal model of the intentionality relation* (PMIR) emerges, by satisfying the constraints of convolved holism, dynamicity, and perspectivalness. PMIR allows one to recognize oneself as a subject (§3.1.5), and allows the creation of an ASM.

The concept of ASM that I developed in §3.2 refers to a collection of mental self-simulations of relations with past and potential future states. The simulational contents are a self in the act of experiencing something that one has experienced before, or that one is likely to experience in future, as well as a temporal and emotional relation between that self and the current experiencing subject—the currently active self-model. We can cognitively and phenomenally access our ASM when it is integrated into our phenomenal self-model. With a phenomenal ASM, one experiences herself as a self living at the present moment and embedded in a

temporal dimension. We have a sense of past experience, events in the past, plans, goals, and expectations for the future. In addition, ASM allows us to own and know our personality, preferences, and values. Differences in ASMs result in differences between individuals.

An ASM is constructed through interaction between episodic and semantic memory systems and the support of PSM and working memory. Three properties are used to characterize ASM: synchronic coherence, diachronic coherence, and global veridicality. Synchronic coherence refers to the consistency of the contents of (self-)simulata of the ASM constructed at a particular time; diachronic coherence refers to consistency between the contents of ASMs constructed at different times; and global veridicality refers to how faithful an ASM simulates past experience and events: the degree of correspondence between the structural and functional features. These properties of ASM are thought to be the constraints of ASM to account for the concerns of authenticity.

Concerning the issue of transtemporal identity, the concepts of transtemporal human identity and transtemporal personal identity are distinguished: The former refers to how a human animal at one time is identical to a human animal at another time; the latter concerns how one person at one time is identical to a person at another time. The biological approach accounts for transtemporal human identity (DeGrazia, 2005b; Olson, 1997, 2003b); however, the second identity relation is based on the concept of personhood. Following the concept of a person suggested by Locke, I hold that personhood is the property of a whole system that emerges when the requirements of membership of the moral community are fulfilled. Thus, personal identity is normative: It relies on how other moral agents regard the identity of other agents.

9.1.4 The Concepts of Authenticity

Debate about authenticity arises because critics and proponents of cognitive enhancement (CE) respectively claim that CE will alienates us and that CE can lead us to an authentic life. “Authenticity”, roughly understood as “being true to one’s self”, is regarded as a value worth pursuing by both camps. However, critics and proponents of CE have comprehended the concept differently, and endorse different ideas of “self”: The former holds a concept of authenticity closer to self-discovery; the latter embraces the concept as a form of self-creation.

One of the main critics of CE, who focuses on concerns about authenticity, is Carl Elliott (1998, 2003). According to Elliott’s ethic of authenticity, life is a

project of one's own in the sense that the meaning and significance of life depend on how one leads one's life project. However, there is no one universal answer to the question of what a good life project is or how to achieve it; each person has to find her own way of living by looking inward: Only through connecting with *our true self* can we *discover* the way in which we can live a good and meaningful life. As such, it is assumed that there exists an internal framework that we ought to discover in order to live authentically.

This concept of authenticity hinges on the idea of self that Elliott endorses. He makes a distinction between "self" (or "the true self") and "self-presentation": The former is what is felt from the inside and is pre-given and constant across time; the latter is what is presented to others and differs from time to time. It is the former that serves as an internal framework to which one ought to conform. To Elliott, the importance of one's identity relies on the maintenance of one's unchangeable true self: Once the true self is altered, one no longer exists as the same person. However, Elliott fails to elucidate what a true self might be. From his illustrative narratives, the concept seems to refer to core psychological traits that determine one's existence. Therefore, for Elliott, what is worrying is not the likelihood of improvement but the fact that it may alter one's true self by changing one's core properties, which are fundamental to one's identity.

Rather than understanding authenticity as self-discovery, David DeGrazia, a proponent of CE, follows the existentialist idea of authenticity, which understands authenticity as self-creation. According to the existentialists, there is no determinant nature of human beings, and what we become is determined by what we do and what we choose. That is, to live authentically, one must do and choose *as myself* instead of as anyone else. DeGrazia (2005a, 2005b), following the idea of authenticity as self-creation, nevertheless holds that we do have an essence, which allows us to persist through time despite some alterations. Unlike Elliott, who regards psychological traits as an essence, DeGrazia adopts animalism and argues that we are essentially human animals.

For DeGrazia, utilizing CE is not morally problematic, because what can be altered by CE is merely one's narrative identity, which refers to psychological characteristics, experiences, or actions attributed to a person, but not to one's essence as a human animal, a change in which would influence one's identity. To be authentic, according to DeGrazia, one has to be honest and autonomous, and an action is autonomous if it is preferred by the subject of the action, if her preference

comes from her identification with the action, and if this identification is not the result of anything that she considers alienating, upon reflection.

In summary, we can identify two conceptions of authenticity: According to authenticity as self-discovery, as endorsed by Elliott, in order to live an authentic life, one has to conform to the internal framework that emerges from one's true self. The true self is pre-given; it remains constant across time, and is constituted by some core psychological traits. On the other hand, authenticity as self-creation, as endorsed by DeGrazia, holds that to act authentically, one has to be honest and autonomous. To be autonomous, one has to prefer and identify with actions through reflection.

9.2 The Phenomenological Account of Memory Enhancement

I argue for the *Phenomenological Account of Enhancement* (PAE) and the *Phenomenological Account of Memory Enhancement* (PAME), which are based on the *Health-Based Account of Enhancement* (HAE) and the *Phenomenological Account of Health* (PAH).

9.2.1 The Phenomenological Concept of Health

According to the HAE, enhancements are distinguished from interventions by the problems to which they respond: Treatments are interventions that address an illness; enhancements are interventions aiming at healthy states or functions. HAE emphasizes that the distinction between treatment and enhancement plays the role of a "moral warning flag" that shows how different interventions should be treated differently. Treatment is more urgent than enhancement, and it is subject to different moral issues: Some normative issues can ineligible for treatment. This account relies heavily on concepts of health; however, different from the disease-based account of enhancement adopted by many, the critical concept is illness rather than disease.

According to Twaddle's (as cited in Hofmann, 2002) distinction between disease and illness, the former refers to the physiological event in a health problem, whereas the latter refers to the subjective undesirable feeling directly perceived by the individual. I agree with Nordenfelt's (2007) Reverse Theory of Disease and Illness which addresses the primacy of illness: The concept of disease is derived from the concept of illness; disease is identified by searching the physiological state which realizes or causes the illness.

PAH is based on the primacy of illness. In contrast to Boorse's (1975, 1977, 1997) objective value-free Biostatistical Theory, according to which the concept of health is normal functioning, which is, in relative to sex, age, and race, the *statistically typical* contribution of all the organism's parts and processes to the organism's goals of its survival and reproduction, PAH does not restrict the goals of the organism to Darwinian values. Rather, PAH is similar to the Holistic Theory of Health (HTH) from Nordenfelt (1993, 2001, 2007), according to which the concept of health is defined by the individual's vital goals—which are necessary and jointly sufficient for his or her minimal long-term happiness.

Both the HTH and the PAH concern an individual's subjective feeling; nevertheless, distinct from HTH, which emphasizes minimal happiness, PAH considers the concept of health free of suffering:

> A subject S is healthy if and only if under standard circumstance, S has the ability to avoid or escape from occurrent or potential suffering.
> A subject S is unhealthy if and only if under standard circumstances, S is currently undergoing suffering and is not able to escape from the situation, or is prone to potential suffering.

Standard circumstance refers to the typical environment in which most individuals in the group live. This concept of health, defined by suffering, builds on the idea of negative utilitarianism, according to which we ought to minimize overall suffering.

9.2.2 The Phenomenological Concept of Enhancement

The Phenomenological Account of Enhancement (PAE) derived from PAH and HAE is as follows:

> *Treatments* are, under standard circumstances, interventions that address the malfunction which results in an individual's suffering or potential suffering that the subject is not able to independently avoid or escape from.
> *Enhancements* are, under standard circumstances and without any unwilling suffering or potential suffering resulted from a target function, interventions that aim at manipulating the target function based on the subject's interests.

In order to examine if an intervention belongs to treatment or enhancement, according to PAE, the following considerations are required.

First, one has to consider if the cognitive function that the treatment addresses is one that results in the suffering of the subject. The mere existence of suffering is not sufficient for any intervention to be considered treatment; instead, only interventions that alleviate suffering by addressing the malfunction are treatments. Therefore, a relationship between cognitive function/malfunction and suffering is required.

Second, to see if an intervention is an enhancement, one not only requires assurance that the suffering—if there is any—results from the cognitive function that the intervention is addressing; one must also investigate one's self-interests. Whether an intervention is considered enhancement or a cognitive function is considered enhanced relies on one's conception, preferences, and values. An intervention is regarded as enhancement only if one identifies and prefers the intervention.

9.2.3 The Phenomenological Concept of Memory Enhancement

To apply PAE to categorize memory interventions:

> *Memory treatments* are, under standard circumstance, interventions that address malfunctions of memory that result in an individual's suffering or potential suffering, and which the subject has no ability to independently escape from.
> *Memory enhancements* are, under standard circumstances and without any unwilling suffering or potential suffering resulting from the alteration of memory functions, interventions that aim to manipulate memory function based on the self-interests of the individual.

As suggested earlier, a relation between cognitive malfunction and suffering is crucial for the phenomenological account of enhancement. Thus, the function and malfunction of memory and how suffering results from its malfunction requires more study. More evidence has shown that the function of memory is to increase individual behavioral flexibility by allowing the (self-)simulation of counterfactual past and future scenarios as well as past events (De Brigard, 2013; Schacter, 2012; Suddendorf & Corballis, 2007). That is, one's misremembering does not necessarily imply the malfunction of memory; instead, one has to see if the functioning of memory successfully contributes to the goals at the higher levels to determine memory malfunction.

But how exactly memory malfunction can result in one's suffering requires more investigation. I have mentioned some ways in which memory malfunction may lead to suffering in §6.2.2, including difficulties in coping with everyday needs, pressure from social interaction and expectations, mood disturbance, and difficulty in maintaining an ASM. The last is noteworthy, as it is related to self-conception and the issue of authenticity.

Failure to satisfy one of the constraints of ASM—synchronic coherence, diachronic coherence, and global veridicality—may result in suffering. For instance, inconsistency between ASMs constructed at different times may results in reluctance to be responsible for one's former decisions or actions, because of a change in one's preferences or goals. A failure to construct an ASM that is globally veridical to a past event may result in difficulties coping with the external world. Nevertheless, a failure to meet the constraints of the diachronic coherence or global veridicality does not necessarily lead to suffering. There are cases in which sacrificing these constraints can promote one's well-being even, for instance, in patients with dementia and dissociative identity disorder. However, one cannot be free from suffering without a synchronically coherent ASM; that is, synchronic incoherence is sufficient for suffering. We rely upon a synchronically coherent ASM to comprehend ourselves and our relations to the external world, including our current and future goals. No matter whether we are healthy or suffering from illness, we strive to construct our ASM with synchronic coherence, either through self-deception or confabulation.

9.3 Memory Enhancement and the Issue of Authenticity

In §7 and §8, I focus on the issue of authenticity and how memory and memory interventions are involved in authenticity. I argue for the following claims:

- Authenticity understood as being true to oneself is conceptually problematic.
- The content of the debate of authenticity can be understood as the constraints of ASM: The concept of authenticity as self-discovery concerns the constraints of diachronic coherence and global veridicality, whereas the concept of authenticity as self-creation concerns the constraints of synchronic coherence and global veridicality.
- The constraint of synchronic coherence of ASM *should* be satisfied,

because synchronic incoherence is sufficient for suffering. Neither the constraint of diachronic coherence nor the constraint of global veridicality is sufficient for the impermissibility of memory interventions.

- The concerns and memory interventions compose a two-way relationship: (1) The concerns understood as the constraints can be used to confine the permissibility of an intervention including memory interventions; (2) memory intervention by modifying the ASM can alter the criteria for the concern of autonomy.
- Autonomy is necessary; however, truthfulness or identity is not sufficient for confining the permissibility of memory intervention or memory treatment; that is, neither concept is sufficient for arguing for or against the permissibility of memory intervention.

9.3.1 The Conceptual Issue of the "True Self"

First, the concept of authenticity is understood as "being true to oneself". According to Elliott's ethics of authenticity, which understands the concept of authenticity as self-discovery, to live authentically one has to discover and conform to a pre-given and static internal framework, which is considered one's "true self" (§7). This is built on the idea that metaphysically there exists a "self" which is determinant of one's existence. However, there is no empirical evidence to support the idea of an ontologically existing thing that is a self.

Erler (2011), for instance, explicitly considers the "true self" a form of narrative identity (Schechtman, 1996). However, acknowledging that narrative identity changes through time, he fails to provide criteria of which psychological characteristics are the core traits that are problematic to alter. In addition, an argument is required for the claim that it would be morally problematic to alter these characteristics. Therefore, to understand the concept of authenticity as being true to one's self is conceptually problematic: One has to explain what the pre-given and static "self" refers to if one does not endorse the ontological realism of the self, and must argue why it *shouldn't* be altered.

9.3.2 Authenticity as Constraint Satisfaction

The issue of authenticity that Elliott and DeGrazia have focused on involves concerns about identity, truthfulness, and autonomy. On the one hand, Elliott argues that the concept of authenticity conforms to the internal framework that he regards

as the "true self". The "true self", according to Elliott, comprises core psychological traits, which determine one's identity. CE is therefore morally problematic, because it might alter these traits and consequently lead to an inauthentic life and loss of identity. Elliott's concept of authenticity involves the issues of identity and truthfulness. On the other hand, DeGrazia defines the concept of authenticity with honesty and autonomy, where autonomy is defined by one's preference and identification. CE is thus not problematic as long as one remains truthful as well as identifying with and preferring CE. DeGrazia's concept of authenticity involves issues of truthfulness and autonomy.

Based on the SMT and the concept of ASM, the issues of identity, autonomy, and truthfulness are respectively analyzed as three constraints of the ASM—diachronic coherence, synchronic coherence, and global veridicality. First, the issue of identity from Elliott concerns whether core psychological characteristics, which determine one's identity, remain the same. This refers to the constraints of diachronic coherence, which relates to the consistency of ASMs constructed at different times. Elliott regards core psychological characteristics as a person's essence, and consequently he endorses the psychological approach of identity. Diachronic coherence—the consistency of the ASMs created at different times—allows one to recognize a previous person as the same person and one's identity relation to hold according to the psychological approach. Next, the concern of autonomy emphasized by DeGrazia and adopted from Frankfurt (1971, 1988), considers how one identifies an action. Autonomy is characterized by the constraint of the synchronic coherence of one's constructed ASM—whether one's action or an intervention can be coherently integrated into one's ASM. For one to identify and prefer an intervention or enhancement after reflection, its application has to be consistent with one's ASM. Last, the issue of truthfulness demanded by both Elliott and DeGrazia refers to the constraint of global veridicality. To be faithful to states of the self and the world, a globally veridical ASM, which represents and simulates current and past states as they are and were experienced in the past, is required.

9.3.3 The Moral Value of the Constraints and Authenticity

To examine if authenticity, understood as two different conceptions by Elliott and DeGrazia, is a moral ideal that is worth pursuing, we can investigate which of these constraints is morally necessary. I argue that the only constraint that we *ought* to satisfy is synchronic coherence, because synchronic incoherence is directly linked to suffering. As mentioned earlier, failure to meet any of these constraints may lead to

suffering; however, only synchronic incoherence is directly linked to suffering. In certain cases in which the three constraints compete with each other, the constraints of diachronic coherence and global veridicality are sacrificed for phenomenal well-being. That is, the constraints of diachronic coherence and global veridicality are not always necessary for one's well-being, and the satisfaction of these constraints can lead to suffering.

This has motivated the reconsideration both of treating the concept of authenticity as a moral ideal as well as the concerns of identity and truthfulness. Commonsensically, we promote these values, for instance the importance of being honest and keeping one's promises. However, if we carefully examine the (self-)simulata generated at different times, we may find them inconsistent and not corresponding to past events. The truth is that we are utilizing the same mechanism dementia patients utilize to confabulate in order to create our ASM based on current goals and circumstances. Furthermore, self-deception is considered a common mechanism in healthy individuals and it has been suggested that this can be beneficial. Thus, whichever conception of authenticity is adopted, we have to respond to the question of why, fundamentally, the loss of identity and truthfulness to the self-conception is morally problematic.

9.3.4 Memory Enhancement and the Constraints

The constraints concern memory interventions in two ways: First, they can be used to confine the permissibility of interventions including memory interventions. Second, memory intervention can alter the ASM to meet or fail the constraints. On the one hand, the concerns of autonomy, identity, and truthfulness, analyzed as constraints of synchronic coherence, diachronic coherence, and global veridicality, are considered by proponents of different concepts of authenticity when thinking about the permissibility of an intervention or enhancement, including memory interventions or enhancements. It is noteworthy that, by definition, concerns about autonomy won't affect memory enhancement, because according to the phenomenological concept of memory enhancement, if one is free from suffering and identifies and prefers the intervention, there won't be synchronic incoherence of ASM.

Unlike interventions or enhancements of other cognitive functions, memory interventions or enhancements are particularly interesting with regard to the issue of authenticity, because memory interventions alter the construction of one's ASM and may result in an alteration of the criteria for the concern of autonomy. That is,

memory interventions can conceptually modify one's ASM to meet or fail the constraint. They may be able to modify one's higher-order volitions (Frankfurt, 1988): For instance, a bioconservativist who wants to become a transhumanist can utilize memory interventions to modify her ASM and thus autonomously utilize CE.

9.4 Open Questions for Future Studies

In the last section, some open questions are presented. The future conceptual and empirical studies on these will either further support the phenomenological account of memory enhancement (§6) and the framework for authenticity (§8) introduced in this dissertation. Based on the conceptual tools developed, the future studies will also provide answers to the normative issues of memory enhancement listed in §1.

1. Everyday memory and the function of memory

Current studies on memory have focused on the mechanisms or neural correlates of individual memory processing systems, such as working, semantic, and episodic memory. However, in order to understand the memory of healthy subjects (as well as memory-related disorders), we need more studies on "everyday memory", which refers to the memory processes that occur in daily environment (G. Cohen, 2008). Different from the memory studies carried out through typical laboratory tasks, how the cooperation of different memory systems results in everyday memory is one of the focuses. There are increasing investigations on how episodic and semantic memory mutually contribute to autobiographical memory and future projections (e.g., Irish & Piguet, 2013); however, a general framework is required. In addition, researches on everyday memory emphasize the fact that everyday memory is context-bound (e.g., social context) and can be externally as well as internally triggered.

Moreover, a new view on the function of memory requires further conceptual and empirical works. This dissertation has adopted the view proposed by De Brigard (2013, see §2.3), according to which the function and malfunction of memory is determined by its contribution to the higher levels. Therefore, in addition to the involvement of memory in future or hypothetical thinking, the relation between memory and other cognitive functions, such as social cognition, future planning, and imagination, requires more attention.

2. The autobiographical self-model

ASM is the conceptual tool proposed in this dissertation to account for how our mental autobiography and related phenomenal experience such as sense of identity are emerged. Based on the ASM, a framework is proposed to provide clearer criteria for the concerns involved in the issue of authenticity (§8). On top of these, this conceptual tool can be useful in accounting for a variety of phenomena in different disciplines ranging from psychiatry (e.g., mental disorders including Capgras syndrome, Fregoli syndrome, and Cotard's syndrome), nursing (e.g., dementia care) to political science (e.g., political attitudes and ideology) and more. In this dissertation (§3.2), I have mainly focused on how memory processing leads to the construction of an ASM. Other cognitive functions including perception, attention, and emotion also play roles in the emergence of an ASM. A better understanding of how their interactions result in an ASM can empower this conceptual tool.

3. The Nature of Suffering and Its relation to the ASM

According to the PAE, the distinction between treatment and enhancement is based on the demarcation of the existence of suffering. Therefore, it is vital to understand the nature of suffering. Suffering has been treated as a phenomenological notion in this dissertation; that is, only systems that have the capacity to experience can suffer (Metzinger, 2013). However, questions remain: What are the minimally sufficient conditions for the emergence of such phenomenal experience? Is there only one kind of phenomenal quality in suffering? Can we quantify suffering or empirically detect the level of suffering? These issues are not only crucial in the distinction between treatment and enhancement, but also allow a more fine-grained empirical study of suffering in patients with mental disorders.

The considerations of the two normative issues—the distinction between memory treatment and enhancement and the issue of authenticity in the context of memory intervention—are built on the significance of one's phenomenal well-being. It is therefore, important to investigate how memory processings directly or indirectly contribute to one's phenomenal well-being or suffering. Based on the answers to the previously proposed questions—the relation of memory and other cognitive functions and the nature of suffering, we may be able to form a causal theory of how suffering can result from memory malfunction, i.e., the failure of memory to contribute to the goals of the higher levels.

4. The concept of authenticity

As we have seen in §7, the problems of the debate of authenticity arise from (1) the different conceptions of authenticity that are endorsed, (2) different accounts of the self and identity embraced, and (3) different imaginations of what enhancement is and could be. I have shown that the concerns of authenticity as self-discovery and as self-creation can be understood as the issues of autonomy, identity and truthfulness, which, based on the framework proposed in §8, can be characterized by the functional constraints of the ASM—synchronic coherence, diachronic coherence, and global veridicality. The framework allows a clear consideration of the concerns in the respects of the influence of cognitive or memory interventions as well as the moral values of these concerns. Therefore, it is worth considering if the term "authenticity" with such ambiguity is worth using. In different contexts, "authenticity" does invovle some worries people may have toward cognitive interventions; however, the use of this term fails to clarify which issue one refers to and which is worth worrying.

Epilogue

Recent research on cognitive enhancement has been focusing on the issues pointing to the utility of the intervention and has overlooked the importance of the phenomenal well-being and self-interests. When considering whether we should modify our cognitive or mental capacitive to gain individual or social advantages, normative issues such as fairness, truthfulness, and moral responsibility are considered, instead of issues regarding our first-person experience. No sufficient attention is devoted to the investigation of the influence that cognitive intervention may bring to our mental life. In addition, the diversity of self-interests is often dismissed. We tend to falsely assume that what the majority prefers equals to a better option. This is not only conceptually problematic but may intensify indirect coercion.

The phenomenological account of (memory) enhancement may not be able to provide an immediate practical criterion to distinguish (memory) enhancement from treatment, because it is built on a conceptual distinction and a theory of suffering of which we are in need. However, this account can be seen as a call to bring the attention of the human enhancement debate to the personal level. Enhancement studies should not only focus on how we can manipulate our biological functioning to gain better utility, but also on how we can phenomenally enhance ourselves. Modern human beings are often troubled by depression, anxiety, stress, etc.; we live with them and suffer from them until they have affected our everyday functioning. This should be considered one of the most important kinds of enhancement.

The autobiographical self-model is one of the central conceptual tools developed in this dissertation. It can be used not only to characterize identity disorders, but also to account for the individual differences and state differences. It shows how different we are from each other, and how careful we should be when attributing normative properties. As we have discussed the interaction normality and normalization, the way we attribute them as well as our conceptions will affect the distribution.

Finally, what I did not address but is fairly important is the influence of social-cultural environment. I have focused on the interventions such as pharmaceuticals and brain stimulation; nevertheless, we may have ignored that manipulating or changing the environment may perhaps be the most effective way.

For instance, a group of patients suffering from Alzheimer's disease in Missouri can live independently because of their specially designed environment (Drayson & Clark, 2007). Likewise, the most powerful cognitive enhancer may be an appropriate modification of the environment of the subject. So far, the most effective memory enhancing technologies are still rather traditional devices, such as notebooks, computers, and smart phones. I believe it is so especially in the issue of moral enhancement.

References

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia, 45*(7), 1363-1377.

Alberini, C. M. (2005). Mechanisms of memory stabilization: Are consolidation and reconsolidation similar or distinct processes? *Trends in Neurosciences, 28*(1), 51-56.

Alberini, C. M. (2009). Transcription factors in long-term memory and synaptic plasticity. *Physiological Reviews, 89*(1), 121-145.

Alberini, C. M. (2011). The role of reconsolidation and the dynamic process of long-term memory formation and storage. *Frontiers in Behavioral Neuroscience, 5*.

Allhoff, F., Lin, P., & Steinberg, J. (2011). Ethics of human enhancement: An executive summary. *Science and Engineering Ethics, 17*(2), 201-212.

Altman, J. (1962). Are new neurons formed in the brains of adult mammals? *Science, 135*(3509), 1127-1128.

American Psychiatric Association. (2013). *The Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396-408.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge, MA: Cambridge University Press.

Baddeley, A. D. (1983). Working memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 302*(1110), 311-324.

Baddeley, A. D. (1992a). What is autobiographical memory? In M. A. Conway, D. C. Rubin, H. Spinnler & W. A. Wagenaar (Eds.), *Theoretical perspectives on autobiographical memory* (Vol. 65, pp. 13-29). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Baddeley, A. D. (1992b). Working memory. *Science, 255*(5044), 556.

Baddeley, A. D. (2000a). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423.

Baddeley, A. D. (2000b). Short-term and working memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 77-92). New York: Oxford University Press.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829-839.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-90). New York: Academic Press.

Baker, L. R. (2000). *Persons and bodies: A constitution view*. Cambridge: Cambridge University Press.

Baker, L. R. (2002). The ontological status of persons. *Philosophy and Phenomenological Research, 65*(2), 370-388.

Baker, L. R. (2013). *Naturalism and the first-person perspective*. Oxford: Oxford University Press.

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1235-1243.

Bar, M., & Neta, M. (2008). The proactive brain: Using rudimentary information to make predictive judgments. *Journal of Consumer Behaviour, 7*(4-5), 319-330.

Barba, G. D., Cappelletti, J. Y., Signorini, M., & Denes, G. (1997). Confabulation: Remembering 'another' past, planning 'another' future. *Neurocase, 3*(6), 425-436.

Barresi, J., & Martin, R. (2011). History as prologue: Western theories of the self. In S. Gallagher (Ed.), *The Oxford handbook of the self*. Oxford: Oxford University Press.

Bartlett, F. C. (1995). *Remembering: An experimental and social psychology*. Cambridge: Cambridge University Press.

Beike, D. R., & Landoll, S. L. (2000). Striving for a consistent life story: Cognitive reactions to autobiographical memories. *Social Cognition, 18*(3), 292-318.

Benatar, D. (2006). *Better never to have been: The harm of coming into existence*. Oxford: Oxford University Press.

Bermúdez, J. L. (2008). Self-consciousness. In M. Velmans & S. Schneider (Eds.), *The blackwell companion to consciousness* (pp. 456-467). Malden, MA: Blackwell Publishing Ltd.

Bernecker, S. (2008). *The metaphysics of memory*. Dordrecht: Springer.

Bernecker, S. (2010). *Memory: A philosophical study*. Oxford: Oxford University Press.

Bernlef, J. (1989). *Out of mind* (A. Dixon, Trans.). Boston, MA: David R. Godine Publisher.

Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science, 4*(4), 370-374.

Birks, J., & Grimley Evans, J. (2009). Ginkgo biloba for cognitive impairment and dementia. *Cochrane Database of Systematic Reviews*.

Boggio, P. S., Fregni, F., Valasek, C., Ellwood, S., Chi, R., Gallate, J., et al. (2009). Temporal lobe cortical electrical stimulation during the encoding and retrieval phase reduces false memories. *PLoS One, 4*(3).

Boorse, C. (1975). On the distinction between disease and illness. *Philosophy & Public Affairs, 5*(1), 49-68.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science, 44*(4), 542-573.

Boorse, C. (1997). A rebuttal on health. In J. M. Humber & R. F. Almeder (Eds.), *What is disease?* Totowa, NJ: Humana Press.

Bostrom, N. (2003). Human genetic enhancements: A transhumanist perspective. *The Journal of Value Inquiry, 37*(4), 493-506.

Bostrom, N., & Roache, R. (2008). Ethical issues in human enhancement. In J. Ryberg, T. Petersen & C. Wolf (Eds.), *New waves in applied ethics* (pp. 120-152). New York: Palgrave Macmillan.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics, 15*(3), 311-341.

Bostrom, N., & Savulescu, J. (2009). *Human enhancement*. Oxford: Oxford University Press.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*(2), 177-220.

Boyer, P. (2008). Evolutionary economics of mental time travel? *Trends in Cognitive Sciences, 12*(6), 219-224.

Boyer, P. (2009). What are memories for? Functions of recall in cognition and culture. In P. Boyer & J. V. Wertsch (Eds.), *Memory in mind and culture* (pp. 3-28). Camnridge: Cambridge University Press.

Breen, N., Caine, D., Coltheart, M., Hendy, J., & Roberts, C. (2000). Towards an understanding of delusions of misidentification: Four case studies. *Mind & Language, 15*(1), 74-110.

Brown, G. P., Macleod, A. K., Tata, P., & Goddard, L. (2002). Worry and the simulation of future outcomes. *Anxiety, Stress & Coping, 15*(1), 1-17.

Bryant, R. A. (2014). Prolonged grief: Where to after Diagnostic and Statistical Manual of Mental Disorders, 5th Edition? *Current Opinion in Psychiatry, 27*(1), 21-26.

Bublitz, J. C. (2013). My mind is mine!? Cognitive liberty as a legal concept. In E. Hildt & A. G. Franke (Eds.), *Cognitive enhancement* (pp. 233-264). Dordrecht: Springer.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*(2), 49-57.

Buss, S. (2013). Personal autonomy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2013 Edition ed.).

Butler, J. (2008). Of personal identity. In J. Perry (Ed.), *Personal identity* (pp. 99-106). Berkeley: University of California Press.

Canguilhem, G. (1978). *On the normal and the pathological*. Dordrecht: D. Reidel Publishing.

Cassell, E. J. (2004). *The nature of suffering and the goals of medicine*. Oxford: Oxford University Press.

Castañeda, H.-N. (2001). 'He': A study in the logic of self-consciousness. In A. Brook & R. C. DeVidi (Eds.), *Self-reference and self-awareness* (Vol. 30, pp. 51-80). Amsterdam: John Benjamins B.V.

Central Intelligence Agency. The world factbook: Definitions and notes. Retrieved November 10, 2013, from https://www.cia.gov/library/publications/the-world-factbook/docs/notesanddefs.html

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19.

Coenen, C., Schuijff, M., Smits, M., Klaassen, P., Hennen, L., Rader, M., et al. (2009). *Human enhancement*. Brussels: STOA, European Parliament.

Cohen, G. (2008). The study of everyday memory. In G. Cohen & M. A. Conway (Eds.), *Memory in the real world* (3rd ed., pp. 1-20). New York: Psychology Press.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.

Conly, S. (2012). *Against autonomy: Justifying coercive paternalism*. Cambridge: Cambridge University Press.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language, 53*(4), 594-628.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review, 107*(2), 261.

Conway, M. A., Singer, J. A., & Tagini, A. (2004). The self and autobiographical memory: Correspondence and coherence. *Social Cognition, 22*(5), 491-529.

Copenhaver, R. (2009). Reid on memory and personal identity. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2009 ed.).

Cortez, M. F. (2012, Febuary 9). Electrical deep-brain stimulation enhances memory in small study. *Bloomberg Bueinessweek*. Retrieved from http://www.businessweek.com/news/2012-02-09/electrical-deep-brain-stimulation-enhances-memory-in-small-study.html

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-114.

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, 53-74.

Crowell, S. (2010). Existentialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2010 ed.).

D'Argembeau, A., & Van der Linden, M. (2006). Individual differences in the phenomenology of mental time travel: The effect of vivid visual imagery and emotion regulation strategies. *Consciousness and Cognition, 15*(2), 342-350.

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.

Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. New York: Pantheon.

Daniels, N. (2000). Normal functioning and the treatment-enhancement distinction. *Cambridge Quarterly of Healthcare Ethics, 9*(3), 309-322.

Davachi, L. (2007). Encoding: The proof is still required. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 137-143). New York: Oxford University Press.

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences, 19*(2), 92-99.

Davis, P. V., & Bradley, J. G. (2000). The meaning of normal. In C. Donley & S. Buckley (Eds.), *What's normal?: Narratives of mental & emotional disorders* (pp. 7-16). Kent, Ohio: The Kent State University Press.

De Brigard, F. (2013). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, 1-31.

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*(12), 2401-2414.

De Jongh, R., Bolt, I., Schermer, M., & Olivier, B. (2008). Botox for the brain: Enhancement of cognition, mood and pro-social behavior and blunting of unwanted memories. *Neuroscience & Biobehavioral Reviews, 32*(4), 760-776.

Debruyne, H., Portzky, M., Van den Eynde, F., & Audenaert, K. (2009). Cotard's syndrome: A review. *Current Psychiatry Reports, 11*(3), 197-202.

DeGrazia, D. (2000). Prozac, enhancement, and self-creation. *The Hastings Center Report, 30*(2), 34-40.

DeGrazia, D. (2005a). Enhancement technologies and human identity. *Journal of Medicine and Philosophy, 30*(3), 261-283.

DeGrazia, D. (2005b). *Human identity and bioethics*. Cambridge, MA: Cambridge University Press.

DeGrazia, D. (2014). What is suffering and what sorts of beings can suffer? In R. M. Green & N. J. Palpant (Eds.), *Suffering and bioethics*. Oxford: Oxford University Press.

DeGrazia, D., & Rowan, A. (1991). Pain, suffering, and anxiety in animals and humans. *Theoretical Medicine and Bioethics, 12*(3).

Dennett, D. C. (1976). Conditions of personhood. In A. Rorty (Ed.), *The identities of persons* (pp. 175-196). Berkeley: University of California Press.

Dennett, D. C. (1991). *Consciousness explained*. New York: Little, Brown and Company.

Dennett, D. C. (1992). The self as a center of narrative gravity. In F. S. Kessel, P. M. Cole & D. L. Johnson (Eds.), *Self and consciousness: Multiple perspectives*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books.

Department of Household Registration. (2012). Statistics: History (End of 2012). Retrieved November 10, 2013, from http://www.ris.gov.tw/en/web/ris3-english/history

Descartes, R. (1985). *The philosophical writings of Descartes: Volume 2* (J. Cottingham, R. Stoothoff & D. Murdoch, Trans.). Cambridge: Cambridge University Press.

Diamond, D. M., Park, C. R., Campbell, A. M., & Woodson, J. C. (2005). Competitive interactions between endogenous LTD and LTP in the hippocampus underlie the storage of emotional memories and stress-induced amnesia. *Hippocampus, 15*(8), 1006-1025.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience, 11*(2), 114-126.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25*(3), 228-245.

Douglas, T. (2013). Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics, 27*(3), 160-168.

Drayson, Z., & Clark, A. (2007). Augmentation, agency, and the spreading of the mental state. *American Journal of Bioethics, 7*(9), 3-11.

Dresler, M., Sandberg, A., Ohla, K., Bublitz, C., Trenado, C., Mroczko-Wąsowicz, A., et al. (2013). Non-pharmacological cognitive enhancement. *Neuropharmacology, 64*, 529-543.

Dresser, R. S. (1981). Ulysses and the psychiatrists: A legal and policy analysis of the voluntary commitment contract. *Harvard Civil Rights-Civil Liberties Law Review, 16*, 777.

Dudai, Y. (2007). Memory: It's all about representations. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 14-16). New York: Oxford University Press.

Dudai, Y., Roediger III, H. L., & Tulving, E. (2007). Memory concepts. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 1-9). New York: Oxford University Press.

Dworkin, G. (1972). Paternalism. *The Monist, 56*(1), 64-84.

Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge: Cambridge University Press.

Dworkin, G. (2010). Paternalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 Edition ed.).

Earp, B. D., Sandberg, A., Kahane, G., & Savulescu, J. (2014). When is diminishment a form of enhancement? Rethinking the enhancement debate in biomedical ethics. *Frontiers in Systems Neuroscience, 8*, 12.

Eichenbaum, H. (2007). Persistence: Necessary, but not sufficient. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 193-197). New York: Oxford University Press.

Elliott, C. (1998). The tyranny of happiness: Ethics and cosmetic psychopharmacology. In E. Parens (Ed.), *Enhancing human traits: Ethical and social implications* (pp. 177-188). Washington, D.C.: Georgetown University Press.

Elliott, C. (1999). *A philosophical disease: Bioethics, culture and identity*. New York, NY: Routledge.

Elliott, C. (2000). Pursued by happiness and beaten senseless: Prozac and the American dream. *Hastings Center Report, 30*(2), 7-12.

Elliott, C. (2003). *Better than well: American medicine meets the American dream*. New York: WW Norton & Company.

Engelhardt, H. T. (1984). Persons and humans: Refashioning ourselves in a better image and likeness. *Zygon, 19*(3), 281-295.

Engelhardt, H. T. (1990). Human nature technologically revisited. *Social Philosophy and Policy, 8*(1), 180-191.

Eriksson, P. S., Perfilieva, E., Björk-Eriksson, T., Alborn, A.-M., Nordborg, C., Peterson, D. A., et al. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine, 4*(11), 1313-1317.

Erler, A. (2011). Does memory modification threaten our authenticity? *Neuroethics, 4*, 235-249.

Eshleman, A. (2009). Moral responsibility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2009 Edition ed.).

Farah, M. J., & Heberlein, A. S. (2007). Personhood and neuroscience: Naturalizing or nihilating? *The American Journal of Bioethics, 7*(1), 37-48.

Feinberg, T. E. (2001). *Altered egos: How the brain creates the self*. New York: Oxford University Press.

Feinberg, T. E. (2005). Neural hierarchies and the self. In T. E. Feinberg & J. P. Keenan (Eds.), *The lost self: Pathologies of the brain and identity* (pp. 33-50). New York: Oxford University Press.

Feinberg, T. E. (2009a). Confabulation, the self, and ego functions: The ego dysequilibrium theory. In W. Hirstein (Ed.), *Confabulation: Views from neuroscience, psychiatry, psychology and philosophy* (pp. 91-107). New York: Oxford University Press.

Feinberg, T. E. (2009b). *From axons to identity: Neurological explorations of the nature of the self*. New York: WW Norton & Company.

Feinberg, T. E. (2011). Neuropathologies of the self: Clinical and anatomical features. *Consciousness and Cognition, 20*(1), 75-81.

Floel, A., & Cohen, L. G. (2007). Contribution of noninvasive cortical stimulation to the study of memory functions. *Brain Research Reviews, 53*(2), 250-259.

Foster, J. (1991). *The immaterial self: A defence of the Cartesian dualist conception of the mind*. London: Routledge.

Franke, A. G., & Lieb, K. (2010). Pharmakologisches Neuroenhancement und „Hirndoping". *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 53*(8), 853-860.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5-20.

Frankfurt, H. G. (1988). *The importance of what we care about: Philosophical essays*. Cambridge: Cambridge University Press.

Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2012). False memories of fabricated political events. *Journal of Experimental Social Psychology*.

Fulford, K. W. M. (1989). *Moral theory and medical practice*. Cambridge: Cambridge University Press.

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences, 4*(1), 14-21.

Glannon, W. (1998). Genes, embryos, and future people. *Bioethics, 12*(3), 187-211.

Glannon, W. (2006). Psychopharmacology and memory. *Journal of Medical Ethics, 32*(2), 74-78.

Glannon, W. (2007). *Bioethics and the brain*. Oxford: Oxford University Press.

Glannon, W. (2011). *Brain, body, and mind: Neuroethics with a human face*. Oxford: Oxford University Press.

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society, 16*(5), 748.

Griffiths, D., Dickinson, A., & Clayton, N. (1999). Episodic memory: What can animals remember about their past? *Trends in Cognitive Sciences, 3*(2), 74-80.

Guan, J.-S., Haggarty, S. J., Giacometti, E., Dannenberg, J.-H., Joseph, N., Gao, J., et al. (2009). HDAC2 negatively regulates memory formation and synaptic plasticity. *Nature, 459*(7243), 55-60.

Guignon, C. B. (2004). Existentialism. Retrieved June 19, 2013, from Routledge: http://www.rep.routledge.com/article/N020

Hackam, D. G. (2007). Translating animal research into clinical benefit. *BMJ: British Medical Journal, 334*(7586), 163.

Hamani, C., McAndrews, M. P., Cohn, M., Oh, M., Zumsteg, D., Shapiro, C. M., et al. (2008). Memory enhancement induced by hypothalamic/fornix deep brain stimulation. *Annals of neurology, 63*(1), 119-123.

Hara, T. (2008). Increasing childlessness in Germany and Japan: Toward a childless society? *International Journal of Japanese Sociology, 17*(1), 42-62.

Harris, C. B., Keil, P. G., Sutton, J., Barnier, A. J., & McIlwain, D. J. (2011). We remember, we forget: Collaborative remembering in older couples. *Discourse Processes, 48*(4), 267-303.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25*(2), 102-111.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences, 104*(5), 1726.

Hasselmo, M. E. (2007). Encoding: Models linking neural mechanisms to behavior. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 123-128). New York: Oxford University Press.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science, 1*(1), 189-202.

Hildt, E. (2013). Cognitive enhancement – A critical look at the recent debate. In E. Hildt & A. G. Franke (Eds.), *Cognitive enhancement* (Vol. 1, pp. 1-14): Springer Netherlands.

Hirst, W., Phelps, E. A., Buckner, R. L., Budson, A. E., Cuc, A., Gabrieli, J. D. E., et al. (2009). Long-term memory for the terrorist attack of September 11: Flashbulb memories, event memories, and the factors that influence their retention. *Journal of Experimental Psychology: General, 138*(2), 161.

Hofmann, B. (2002). On the triad disease, illness and sickness. *Journal of Medicine and Philosophy, 27*(6), 651 – 673.

Howard Hughes Medical Institute. (2009, May 7). Relax and learn: New drugs that help DNA unwind may improve memory. Retrieved Febuary 27, 2013, from http://www.hhmi.org/news/tsai20090507.html

Hume, D. (2004). *A treatise of human nature*. Lawrence: Digireads.com Publishing.

Humphrey, N., & Dennett, D. C. (1989). Speaking for our selves: An assessment of multiple personality disorder. *Raritan, 9*(1), 68-98.

Hyman Jr, I. E., & Loftus, E. F. (1998). Errors in autobiographical memory. *Clinical Psychology Review, 18*(8), 933-947.

Irish, M., & Piguet, O. (2013). The pivotal role of semantic memory in remembering the past and imagining the future. *Frontiers in Behavioral Neuroscience, 7*(27).

James, W. (1890). *The principles of psychology*. New York: Holt, Rinehart & Winston.

Jay, T. M. (2003). Dopamine: A potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology, 69*(6), 375-390.

Jetten, J., Haslam, C., Pugliese, C., Tonks, J., & Haslam, S. A. (2010). Declining autobiographical memory and the loss of identity: Effects on well-being. *Journal of Clinical and Experimental Neuropsychology, 32*(4), 408-416.

Jones, D. G. (2006). Enhancement: Are ethicists excessively influenced by baseless speculations? *Medical Humanities, 32*(2), 77-81.

Juengst, E. T. (1998). What does enhancement mean? In E. Parens (Ed.), *Enhancing human traits: Ethical and social implications* (pp. 29-47). Washington, D.C.: Georgetown University Press.

Kadlec, E. (2008). Popper's "negative utilitarianism": From utopia to reality. In P. K. Markl, E. (Ed.), *Karl Popper's response to 1938* (pp. 107-121). Frankfurt am Main: Peter Lang GmbH.

Kandel, E. R. (2007). *In search of memory: The emergence of a new science of mind*. New York: W. W. Norton & Company.

Kays, J. L., Hurley, R. A., & Taber, K. H. (2012). The dynamic brain: Neuroplasticity and mental health. *J Neuropsychiatry Clin Neurosci, 24*(2), 118-124.

Kitcher, P. (1997). *The lives to come: The genetic revolution and human possibilities* New York: Free Press.

Klein, S. B., Cosmides, L., Gangi, C. E., Jackson, B., Tooby, J., & Costabile, K. A. (2009). Evolution and episodic memory: An analysis and demonstration of a social function of episodic recollection. *Social Cognition, 27*(2), 283.

Klein, S. B., Loftus, J., & Kihlstrom, J. F. (2002). Memory and temporal experience: The effects of episodic memory loss on an amnesic patient's

ability to remember the past and imagine the future. *Social Cognition, 20*(5), 353-379.

Klein, S. B., & Nichols, S. (2012). Memory and the sense of personal identity. *Mind, 121*(483), 677-702.

Kopelman, M., Wilson, B., & Baddeley, A. (1989). The autobiographical memory interview: A new assessment of autobiographical and personal semantic memory in amnesic patients. *Journal of Clinical and Experimental Neuropsychology, 11*(5), 724-744.

Kovács, J. (1998). The concept of health and disease. *Medicine, Healthcare and Philosophy, 1*(1), 31-39.

Kraemer, F. (2011). Me, myself and my brain implant: Deep brain stimulation raises questions of personal authenticity and alienation. *Neuroethics*, 1-15.

Lacy, J. W., & Stark, C. E. (2013). The neuroscience of memory: Implications for the courtroom. *Nature Reviews Neuroscience, 14*(9), 649-658.

Lawlor, K. (2009). Memory. In B. McLaughlin, A. Beckermann & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 663-677). Oxford: Oxford University Press.

Lee, Y. S., & Silva, A. J. (2009). The molecular and cellular biology of enhanced cognition. *Nature Reviews Neuroscience, 10*(2), 126-140.

Liao, S. M., & Sandberg, A. (2008). The normativity of memory modification. *Neuroethics, 1*(2), 85-99.

Lin, Y.-T. (2015). Memory for prediction error minimization: From depersonalization to the delusion of non-existence—A commentary on Philip Gerrans. In T. Metzinger & J. M. Windt (Eds.), *Open Mind*. Frankfurt am Main: Mind Group.

Lisman, J., & Redish, A. D. (2009). Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1193-1201.

Locke, J. (2008). *An essay concerning human understanding*. Oxford: Oxford University Press.

Loftus, E. F. (1997). Creating childhood memories. *Applied Cognitive Psychology, 11*(7), S75-S86.

Luria, A. R. (1968). *The mind of a mnemonist: A little book about a vast memory*. New York: Basic Books, Inc.

Lynch, G. (2002). Memory enhancement: The search for mechanism-based drugs. *Nature Neuroscience, 5*, 1035-1038.

Lynch, G. (2004). AMPA receptor modulators as cognitive enhancers. *Current Opinion in Pharmacology, 4*(1), 4-11.

Lynch, G., & Gall, C. M. (2006). Ampakines and the threefold path to cognitive enhancement. *Trends in Neurosciences, 29*(10), 554-562.

Maguire, E. A., Kumaran, D., Hassabis, D., & Kopelman, M. D. (2010). Autobiographical memory in semantic dementia: A longitudinal fMRI study. *Neuropsychologia, 48*(1), 123-136.

Malenka, R. C. (2002). Synaptc pasticity. In K. L. Davis, D. Charney, J. T. Coyle & C. Nemeroff (Eds.), *Neuropsychopharmacology: The fifth generation of progress* (pp. 147-158). Philadelphia: Lippencott Williams & Wilkins.

Marshall, E. (2004). A star-studded search for memory-enhancing drugs. *Science, 304*, 36-38.

Martin, C. B., & Deutscher, M. (1966). Remembering. *The Philosophical Review, 75*(2), 161-196.

Mateer, C. A., & Bogod, N. M. (2003). Cognitive, behavioral, and selected pharmacologic interventions in rehabilitation after acquired brain injury. In R. B. Schiffer, S. M. Rao & B. S. Fogel (Eds.), *Neuropsychiatry* (2nd ed., pp. 165-186). Philadelphia: Lippincott Williams & Wilkins.

Mayerfeld, J. (1999). *Suffering and moral responsibility*. New York: Oxford University Press.

McCabe, S. E., Knight, J. R., Teter, C. J., & Wechsler, H. (2005). Non medical use of prescription stimulants among US college students: Prevalence and correlates from a national survey. *Addiction, 100*(1), 96-106.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*(2), 159-188.

McGinn, C. (2004). *Mindsight: Image, dream, meaning*. Cambridge, MA: Harvard University Press.

Menary, R. (2010). *The extended mind*. Cambridge: The MIT Press.

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences, 2*(4), 353-393.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: The MIT Press.

Metzinger, T. (2005). Précis: Being no one. *Psyche, 11*(5), 1-35.

Metzinger, T. (2007). Self models. *Scholarpedia, 2*, 4174.

Metzinger, T. (2008). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. In R. Banerjee & B. K. Chakrabarti (Eds.), *Models of brain and mind: Physical, computational, and psychological approaches* (pp. 215-245). Amsterdam, Netherlands: Elsevier.

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. New York: Basic Books.

Metzinger, T. (2011). The no-self alternative. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 279-296). Oxford: Oxford University Press.

Metzinger, T. (2013). Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.), *Robotik und Gesetzgebung*. Baden-Baden: Nomos.

Metzinger, T., & Hildt, E. (2011). Cognitive enhancement. In J. Illes & B. J. Sahakian (Eds.), *The Oxford handbook of neuroethics* (pp. 245-264). Oxford: Oxford University Press.

Michaelian, K. (2012). Is external memory memory? Biological memory and extended mind. *Consciousness and Cognition, 21*(3), 1154-1165.

Migues, P. V., Hardt, O., Wu, D. C., Gamache, K., Sacktor, T. C., Wang, Y. T., et al. (2010). PKMzeta maintains memories by regulating GluR2-dependent AMPA receptor trafficking. *Nature Neuroscience, 13*(5), 630-634.

Mill, J. S. (2004). *Principles of political economy: With some of their applications to social philosophy*. Indianapolis: Hackett Publishing Company.

Mill, J. S. (2009). *On liberty*. Auckland: The Floating Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, MA: MIT Press.

Milner, B. (1966). Amnesia following operation on the temporal lobes. In C. W. M. Whitty & O. L. Zangwill (Eds.), *Amnesia* (pp. 109-133). London: Butterworths.

Minsky, M. L. (2006). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. New York: Simon & Schuster.

Morris, R. G. M. (2007). Memory: Distinctions and dilemmas. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 29-34). New York: Oxford University Press.

Moscovitch, M. (2007). Memory: Why the engram is elusive. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 17-21). New York: Oxford University Press.

Moscovitch, M., Chein, J. M., Talmi, D., & Cohn, M. (2010). Learning and memory. In B. J. Baars & N. M. Gage (Eds.), *Cognition, brain, and consciousness: Introduction to cognitive neuroscience*. London, UK: Academic Press.

Murphy, D. (2009). Concepts of disease and health. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2009 ed.).

Murphy, E. A. (1966). A scientific viewpoint on normalcy. *Perspectives in Biology and Medicine, 9*(3), 333-348.

Nader, K., & Einarsson, E. Ö. (2010). Memory reconsolidation: An update. *Annals of the New York Academy of Sciences, 1191*(1), 27-41.

Nava, E., & Röder, B. (2011). Adaption and maladaptation: Insights from brain plasticity. In A. M. Green, C. E. Chapman, J. F. Kalaska & F. Lepore (Eds.), *Enhancing performance for action and perception: Multisensory integration, neuroplasticity and neuroprosthetics* (pp. 177-194). Amsterdam: Elsevier.

Nordenfelt, L. (1993). Concepts of health and their consequences for health care. *Theoretical Medicine, 14*(4), 277-285.

Nordenfelt, L. (2001). *Health, science, and ordinary language*. Amsterdam: Rodopi Publishers.

Nordenfelt, L. (2007). The concepts of health and illness revisited. *Medicine, Health Care and Philosophy, 10*(1), 5-10.

Olson, E. T. (1997). *The human animal: Personal identity without psychology*. Oxford: Oxford University Press.

Olson, E. T. (2003a). An argument for animalism. In R. Martin & J. Barresi (Eds.), *Personal identity* (pp. 318-334). Oxford: Blackwell Publishing.

Olson, E. T. (2003b). Personal identity. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell Guide to Philosophy of Mind* (pp. 352-368). Malden, MA: Blackwell Publishing.

Parens, E. (2005). Authenticity and ambivalence: Toward understanding the enhancement debate. *Hastings Center Report, 35*(3), 34-41.

Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.

Parsons, C. G., Stöffler, A., & Danysz, W. (2007). Memantine: A NMDA receptor antagonist that improves memory by restoration of homeostasis in the glutamatergic system - too little activation is bad, too much is even worse. *Neuropharmacology, 53*(6), 699-723.

PeakNootropics. (2013, Febuary 3). New nootropic drugs—meet the ampakines. Retrieved March 1, 2013, from http://peaknootropics.com/new-nootropic-drugs/

Peters, J., & Büchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron, 66*(1), 138-148.

Pogačić Kramp, V., & Herrling, P. (2011). List of Drugs in Development for Neurodegenerative Diseases: Update June 2010. *Neurodegenerative Diseases, 8*(1-2), 44-94.

Popper, K. R. (1971). *The open society and its enemies, Volume I*. London: Princeton University Press.

Poreisz, C., Boros, K., Antal, A., & Paulus, W. (2007). Safety aspects of transcranial direct current stimulation concerning healthy subjects and patients. *Brain Research Bulletin, 72*(4–6), 208-214.

President's Council on Bioethics. (2003). *Beyond therapy: Biotechnology and the pursuit of happiness*. Retrieved from http://www.bioethics.gov/reports/beyondtherapy/.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(2), 676-682.

Ramachandran, V. S. (2003). *The emerging mind*. London: Profile Books.

Reid, T. (1769). *An inquiry into the human mind on the principles of common sense*. London: Printed for T. Cadell.

Reid, T. (1786). *Essays on the intellectual powers of man*. Dublin: Printed for L. White.

Reid, T. (1878). *Essays on the intellectual powers of man*. Philadelphia: J. H. Butler & Co.

Reid, T. (2008). Of identity. In J. Perry (Ed.), *Personal identity* (pp. 107-112). Berkeley: University of California Press.

Rubin, D. (1986). Introduction. In D. Rubin (Ed.), *Autobiographical memory* (pp. 3-16). Cambridge, UK: Cambridge University Press.

Rubin, D. (1999). Introduction. In D. Rubin (Ed.), *Remembering our past: Studies in autobiographical memory* (pp. 1-15). Cambridge, UK: Cambridge University Press.

Ruby, F. J., Smallwood, J., Sackur, J., & Singer, T. (2013). Is self-generated thought a means of social problem solving? *Frontiers in Psychology, 4*, 1-10.

Russell, B. (2009). *The analysis of mind*. Auckland: The Floating Press.

Sabin, J. E., & Daniels, N. (1994). Determining "medical necessity" in mental health practice. *Hastings Center Report, 24*(6), 5-13.

Sams, M., Hari, R., Rif, J., & Knuutila, J. (1993). The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence. *Journal of Cognitive Neuroscience, 5*(3), 363-370.

Savulescu, J., Sandberg, A., & Kahane, G. (2011). Well-being and enhancement. In J. Savulescu, R. ter Meulen & G. Kahane (Eds.), *Enhancing human capacities* (pp. 3-18). Oxford: Blackwell Publishing.

Savulescu, J., ter Meulen, R., & Kahane, G. (2011). *Enhancing human capacities*. Oxford: Wiley-Blackwell.

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers*. MA: Houghton Mifflin.

Schacter, D. L. (2007). Memory: Delineating the core. In H. L. Roediger III, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 23-27). New York: Oxford University Press.

Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *American Psychologist, 67*(8), 603.

Schacter, D. L., & Addis, D. R. (2007a). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1481), 773-786.

Schacter, D. L., & Addis, D. R. (2007b). Constructive memory: The ghosts of past and future. *Nature, 445*(7123), 27-27.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic simulation of future events. *Annals of the New York Academy of Sciences, 1124*(1), 39-60.

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences, 15*(10), 467-474.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49*(1), 289-318.

Schacter, D. L., Verfaellie, M., & Pradere, D. (1996). The neuropsychology of memory illusions: False recall and recognition in amnesic patients. *Journal of memory and language, 35*(2), 319-334.

Schechtman, M. (1996). *The constitution of selves*. New York: Cornell University Press.

Schermer, M. (2013). Health, happiness and human enhancement—Dealing with unexpected effects of deep brain stimulation. *Neuroethics, 6*(3), 435-445.

Schneider, S. (2009). Mindscan: Transcending and enhancing the human brain. In S. Schneider (Ed.), *Science fiction and philosophy: From time travel to superintelligence* (pp. 241). Malden, MA: Blackwell Publishing.

Schüpbach, M., Gargiulo, M., Welter, M., Mallet, L., Behar, C., Houeto, J., et al. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*(12), 1811-1816.

Seager, W. E., & Bourget, D. (2007). Representationalism about consciousness. In M. Velmans & S. Schneider (Eds.), *The blackwell companion to consciousness* (pp. 261-276). Malden, MA: Blackwell Publishing Ltd.

Seigel, J. E. (2005). *The idea of the self: Thought and experience in Western Europe since the seventeenth century*. Cambridge, MA: Cambridge University Press.

Senor, T. D. (2009). Epistemological problems of memory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2009 ed.).

Sententia, W. (2004). Neuroethical considerations: Cognitive liberty and converging technologies for improving human cognition. *Annals of the New York Academy of Sciences, 1013*(1), 221-228.

Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology, 22*(2), 261-273.

Shema, R., Sacktor, T. C., & Dudai, Y. (2007). Rapid erasure of long-term memory associations in the cortex by an inhibitor of PKM zeta. *Science, 317*(5840), 951-953.

Shoemaker, S. (1970). Persons and their pasts. *American Philosophical Quarterly, 7*(4), 269-285.

Siderits, M. (2011). Buddhist non-self: The no-owner's manual. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 297-315). Oxford: Oxford University Press.

Singer, P. (1997). The drowning child and the expanding circle. *New Internationalist, 289,* 28-30.

Singer, P. (2011). *Practical ethics*. Cambridge, MA: Cambridge University Press.

Singh, I. (2012). Not robots: Children's perspectives on authenticity, moral agency and stimulant drug treatments. *Journal of Medical Ethics, 39*(6), 359-366.

Smallwood, J., Ruby, F. J. M., & Singer, T. (2013). Letting go of the present: Mind-wandering is associated with reduced delay discounting. *Consciousness and Cognition, 22*(1), 1-7.

Smith, G. S., Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Diaconescu, A. O., Workman, C. I., et al. (2012). Increased cerebral metabolism after 1 year of deep brain stimulation in Alzheimer disease. *Archives of Neurology, 69*(9), 1141-1148.

Smith, M. A., Riby, L. M., Eekelen, J. A. M. v., & Foster, J. K. (2011). Glucose enhancement of human memory: A comprehensive research review of the glucose memory facilitation effect. *Neuroscience & Biobehavioral Reviews, 35*(3), 770-783.

Snowdon, P. F. (1991). Personal identity and brain transplants. In D. Cockburn (Ed.), *Human beings* (pp. 109-126). Cambridge: Cambridge University Press.

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489-510.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*(3), 171-177.

Squire, L. R., & Kandel, E. (2009). *Memory: From mind to molecules* (2nd ed.). Greenwood Village: Roberts & Company Publishers.

Stern, S. A., & Alberini, C. M. (2013). Mechanisms of memory enhancement. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 5*(1), 37-53.

Strawson, G. (1999a). The self. In S. Gallagher & J. Shear (Eds.), *Models of the Self* (pp. 1-24). Exeter: Imprint Academic.

Strawson, G. (1999b). The self and the SESMET. In S. Gallagher & J. Shear (Eds.), *Models of the Self* (pp. 483-518). Exeter: Imprint Academic.

Suddendorf, T., & Busby, J. (2003). Mental time travel in animals? *Trends in Cognitive Sciences, 7*(9), 391-396.

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs, 123*(2), 133-167.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences, 30*(03), 299-313.

Suthana, N., Haneef, Z., Stern, J., Mukamel, R., Behnke, E., Knowlton, B., et al. (2012). Memory enhancement and deep-brain stimulation of the entorhinal area. *New England Journal of Medicine, 366*(6), 502-510.

Sutton, J. (1998). *Philosophy and memory traces: Descartes to connectionism*. Cambridge, MA: Cambridge University Press.

Sutton, J. (2010). Memory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 ed.).

Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences, 9*(4), 521-560.

Szpunar, K. K., Watson, J. M., & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences, 104*(2), 642-647.

Taylor, C. (1989). *Sources of the self: The making of the modern identity*. Cambridge: Harvard University Press.

Taylor, C. (1991). *The Ethics of authenticity*. Cambridge, MA: Cambridge University Press.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193-210.

Tchantchou, F., Xu, Y., Wu, Y., Christen, Y., & Luo, Y. (2007). EGb 761 enhances adult hippocampal neurogenesis and phosphorylation of CREB in transgenic mouse model of Alzheimer's disease. *The FASEB Journal, 21*(10), 2400-2408.

Teter, C. J., McCabe, S. E., LaGrange, K., Cranford, J. A., & Boyd, C. J. (2006). Illicit use of specific prescription stimulants among college students: Prevalence, motives, and routes of administration. *Pharmacotherapy, 26*(10), 1501.

Tully, T., Bourtchouladze, R., Scott, R., & Tallman, J. (2003). Targeting the CREB pathway for memory enhancers. *Nature Reviews Drug Discovery, 2*(4), 267-277.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381-402). New York: Academic Press.

Tulving, E. (1983). Elements of episodic memory. Oxford: Clarendon Press.

Tulving, E. (1985a). How many memory systems are there? *American Psychologist, 40*(4), 385.

Tulving, E. (1985b). Memory and consciousness. *Canadian Psychology, 26*(1), 1-12.

Tulving, E. (1987). Multiple memory systems and consciousness. *Human Neurobiology, 6*(2), 67-80.

Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science, 2*(3), 67-70.

Tulving, E. (1995). Organization of memory: Quo vadis. *The Cognitive Neurosciences*, 839-847.

Tulving, E. (2000). Concepts of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 33-43). New York: Oxford University Press.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human. In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3-56). Oxford: Oxford University Press.

Tye, M. (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.

United Nations. (2001). *World population ageing: 1950-2050*. New York.

van Roojen, M. (2012). Moral cognitivism vs. non-cognitivism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34*, 1-56.

Vosgerau, G. (2010). Memory and content. *Consciousness and Cognition, 19*(3), 838-846.

Walters, L., & Palmer, J. G. (1996). *The ethics of human gene therapy*. New York: Oxford University Press.

Wells, G. L., & Loftus, E. F. (2003). Eyewitness memory for people and events. *Handbook of psychology*.

Westbury, C., & Dennett, D. C. (2000). Mining the past to construct the future: Memory and belief as forms of knowledge. In D. L. Schacter & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 11-32). Cambridge, MA: Harvard University Press.

WHO consultation. (2000). Obesity: Preventing and managing the global epidemic. *World Health Organization Technical Report Series, 894*.

WHO expert consultation. (2004). Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *The Lancet, 363*(9403), 157-163.

Williams, H., Conway, M. A., & Cohen, G. (2008). Autobiographical memory. In G. Cohen & M. A. Conway (Eds.), *Memory in the real world* (3rd ed., pp. 21-90). New York: Psychology Press.

Zehetleitner, M., & Schönbrodt, F. D. (2013). When misrepresentation is successful. In T. Breyer (Ed.), *Epistemological foundations of evolutionary psychology.* New York: Springer.

Zusammenfassung

Das wachsende Verständnis von menschlicher Kognition und die Entwicklung neuro-technologischer Fertigkeiten haben eine Debatte über künstlich herbeigeführte Verbesserung kognitiver Funktionen (*Enhancement*) im Bereich der Neuroethik hervorgebracht. Diese Doktorarbeit hat das Ziel, die normativen Implikationen des *Enhancement* im Bereich des Gedächtnisses zu untersuchen und legt dabei die folgenden inhaltlichen Schwerpunkte: (1) die Unterscheidung zwischen Behandlung und *Enhancement* von Gedächtnisleistungen; und (2) der Zusammenhang zwischen dem Thema Authentizität und Eingriffen in Gedächtnisleistungen.

Der erste Teil der Arbeit enthält eine begriffliche Analyse der für die normativen Überlegungen relevanten Terminologie. Dabei werden zunächst die repräsentationale Struktur und die Funktion des Gedächtnisses diskutiert. Gedächtnisleistungen werden als spezielle Formen der Selbst-Repräsentation beschrieben, welche aus einem konstruktiven Prozess resultieren. Darauffolgend untersuche ich die Begriffe Selbst, Person(-haftigkeit) und Identität, wobei das „autobiographische Selbstmodell“ (ASM) als begriffliches Werkzeug eingeführt wird. Ein ASM wird angesehen als eine Sammlung mentaler Repräsentationen der Relationen zwischen zukünftigen und vergangenen Systemzuständen. Ferner soll die Debatte um objektivistische bzw. konstruktivistische Ansichten über Gesundheit in Augenschein genommen werden. Ich argumentiere hier für eine phänomenologische Auffassung von Gesundheit, welche auf dem Primat der Krankheit und dem negativen Utilitarismus basiert.

Der zweite Teil der Dissertation führt die erarbeiteten begrifflichen Werkzeuge mit normativen Überlegungen zusammen. Mein Argument stützt eine auf Leiden basierte Unterscheidung zwischen Behandlung und *Enhancement*. Letzteres wird dabei als eine Intervention betrachtet, welche gezielt Gedächtnisfunktionen im Interesse des Individuums manipuliert. Dies geschieht unter normalen Umständen und ohne Verbindung zu einem Leiden, welches mit Veränderungen von Gedächtnisleistungen einhergeht. Als nächstes wird unter dem Aspekt der Authentizitätsdebatte überlegt, ob Interventionen und *Enhancement* ‚das wahre Selbst‘ gefährden. Ich diskutiere dabei zwei Auslegungen des Begriffs: Authentizität als Selbstentdeckung und Authentizität als Selbstkonstruktion. Während dieser begrifflichen Analyse schlage ich vor, dass das Problem der

Authentizität am besten beschrieben werden kann, wenn es im Rahmen von Erfüllungen funktionaler Bedingungen des ASM betrachtet wird. Diese sind die synchrone und diachrone Kohärenz sowie globale Wahrhaftigkeit.

Der nun erarbeitete begriffliche Rahmen bietet klarere Kriterien, unter welchen die relevanten Probleme betrachtet werden können, und ermöglicht so eine Untersuchung moralischer Aspekte von Authentizität.