Introduction

In classical mythology, the souls of the dead who drank from the river Mnemosyne would remember everything from their experience. Those who drank from the river Lethe would forget everything and enter the realm of oblivion. In our actual lives, memory capacity and impairment fall along a spectrum between these two extremes. Unlike the characters in mythology, we cannot choose how much or how little memory we have. Working memory, retrieval of episodic and semantic memory and the initial learning in procedural memory are to some extent within our conscious control. But we have no control over the encoding, consolidation, storage and reconsolidation of memory. This may change, however, with interventions designed to increase memory capacity, alter the content of memories or erase them.

Memory is a vital process in humans. At the most basic biological level, the capacity of the adaptive arm of the immune system to form a memory of antigens enables it to recognize and eliminate pathogens through the combined action of antibodies, complement and macrophages. Antigenic memory is thus necessary for the survival of the organism. At more evolved neurobiological and psychological levels, learning mediated by subcortical brain structures enables us to perform motor skills automatically without having to think about performing them. Brainstem structures responding to sensory stimuli send inputs to the hippocampus that allow us to remember new places. Memories of threatening events mediated by the brain's fear memory system allow us to recognize new threats and confront or avoid them. These are further examples of how memory is critical to our survival. At a psychological level, the experience of mental time travel in recalling the past and imagining the future gives one the feeling of persisting through time as the same person. It allows one to integrate one's experiences into a coherent whole and construct a meaningful autobiography. Information about the past enables us to engage in goal-directed behavior in forming and executing action plans.

Robert Veselis emphasizes the significance of the memory process: "Memory makes us uniquely human. As the human mind is the most

complex creation in the universe, it stands to reason that memory embodies to a large extent this complexity. When memory fails in the end for some of us, a large portion of our being human also fails" (Veselis, 2017, p. 31).

Visual, auditory, gustatory and olfactory cues can trigger autobiographical memories transporting us back to childhood or places where we experienced certain sights, sounds, tastes and smells. Memories of the departed allow them to visit us in dreams. They may console or haunt us. The Ghost in *Hamlet* is the mental representation of the main character's father. In the same play, the graveyard scene in which Hamlet reflects on the deceased court jester Yorick when his skull is exhumed is another example of the power of episodic memory. Recall of past misdeeds or omissions can generate regret and other emotions that can influence our current and future behavior in beneficial or harmful ways. Memories others have of us may provide a sense of virtual survival beyond death. But all of these memories eventually dissolve in oblivion.

Brain injury and neurological disorders can disrupt the brain's capacity to encode, consolidate, store, retrieve and reconsolidate memories. This disruption can adversely affect the psychological capacities associated with memory and have a deleterious effect on people's lives. The inability to form new memories or retrieve existing memories can impair or undermine the experience of persisting through time and the capacity for agency. In other circumstances, an emotionally charged memory of a traumatic experience may become firmly entrenched in the brain and mind and cause the psychopathology characteristic of posttraumatic stress disorder (PTSD), panic and anxiety disorders. Depending on how it affects our thought and behavior, memory can have value or disvalue for us.

This book is a thematically integrated analysis and discussion of neuroethical questions about memory and interventions to modify it. It is written for a multidisciplinary audience, including psychologists, clinical neuroscientists, philosophers, bioethicists, legal theorists and informed lay readers. By discussing historical and current theories of memory, and examining existing and emerging forms of memory modification, the book shows how empirical and normative aspects of memory have evolved. The subtitle of the book captures the spectrum of memory extending from exceptional recall to profound amnesia and advanced dementia. These opposite ends of the spectrum are the rough equivalents of the mythological Mnemosyne and Lethe. Our ability to adapt to changing environments requires optimal levels of memory capacity and content between these extremes. A certain amount of information about the past is necessary to plan and act in the present and future. But too much of this information can overload the brain and mind and interfere

with reasoning and decision-making. Some degree of forgetting is necessary to learn new information relevant to one's natural and social milieu. Flexible thought and behavior require a balance between remembering and forgetting.

Although different memory systems support different physical and mental functions, I focus mainly on two subtypes of declarative memory: episodic and semantic. Episodic memory is knowledge of events that happened at a specific time and place (Tulving, 1983). Semantic memory is knowledge of facts and concepts about the world (Tulving, 1985a). Working and prospective memory are necessary for rational and moral agency. These two declarative memory subtypes rely on episodic and semantic memory, which have a broader range of functions. Episodic memory is necessary not only for agency but also for identity. Among memory systems, episodic and semantic memory are most pertinent to metaphysical, ethical and legal questions about identity, agency, responsibility, benefit and harm.

More specifically, how do normal memory functions enable us to initiate and execute action plans? How does memory dysfunction impair this ability? To what extent is personal identity based on the capacity to accurately recall the past? How many memories could be lost without causing a substantial change in identity? Could a person with early-stage dementia exercise precedent autonomy in expressing earlier wishes about later life-sustaining care when she has no memory of these wishes? If a patient under general anesthesia becomes aware intraoperatively, then would it be permissible for an anesthetist to infuse an amnesia-inducing drug without the patient's prior consent? Would the patient be harmed if she had no memory of being aware? How would the patient know that she was aware without a memory of it? How do we weigh the potential neurological and psychological benefit against the risk of harm from brain implants designed to improve or restore some memories and weaken or erase others? Do we discover our true selves through the backward-looking aspects of memory or create them through the forward-looking aspects of memory? How would modifying memories influence authenticity? Can a person be responsible for an action if she does not remember performing it? Would a victim of a criminal act have to duty to retain a memory of it to testify against the perpetrator? Or would her cognitive liberty give her the right to erase the memory? Focusing mainly on disorders of memory content and capacity, I use actual and hypothetical cases to analyze and discuss these questions.

What it is like to recall an experience is more than a function of neurobiology. Still, we cannot understand memory without understanding its neurobiological underpinning. Interactions between cortical and

subcortical brain regions allow the integration of information about a person's experience into a coherent and consciously accessible representation of it. Although our experience as rational and moral agents and subjects persisting through time is a psychological property, it is possible through the normal function of neurobiological processes that enable memory. Neuroscientific and behavioral research has helped to explain normal memory function, as well as how transient and chronic neurological disorders can result in different types of memory dysfunction. This research on memory and how it influences our thought and behavior forms the theoretical basis of this book.

While I explore the philosophical *implications* of memory, I do not engage with philosophical *theories* of memory (Sutton, 1998; Bernecker, 2008, 2010; Bernecker and Michaelian, 2017). Many of these theories do not adequately consider the multiple neural networks that mediate different memory functions. Some ignore these networks altogether and discuss memory exclusively in psychological terms. Yet failing to account for both psychological and neurobiological aspects of memory results in what is at best an incomplete explanation of the different systems, types and subtypes of memory. In normal circumstances, whether a person remembers events or facts can be confirmed by verbal reports and her general behavior. Yet memory disorders have provided the best evidence of memory and its role in our conscious and unconscious life. These usually result from brain damage and dysfunction. Some forms of amnesia may be psychogenic, though these tend to be transient. Dissociative disorders are typically described as psychogenic. But they may correlate with detectable neurobiological changes. Although they manifest in varying types and degrees of mental impairment, most memory disorders are associated with anatomical and functional abnormalities in the brain.

Some philosophers distinguish between experiential, propositional (or factual) and practical memory (Bernecker, 2008, 2010). This taxonomy is consistent with psychologists' and cognitive neuroscientists' distinction between episodic, semantic and procedural memory (Tulving, 1983, 1985a). Philosophers tend to explain memory content in terms of experiential (nonpropositional) or factual (propositional) attitudes. The first type of attitude has a first-person content, and the second type of attitude has a third-person content (Burge, 2003). These attitudes correspond roughly to autobiographical and semantic memory. Yet the abstract and at times overly technical formulation of them fails to show how they manifest in actual thought and behavior.

Distinguishing between direct realist and representational theories of memory, some philosophers discuss whether we have direct or indirect

access to past events (Sutton 1998; Bernecker, 2008, 2010). According to direct realism, when we recall an event, we are in direct cognitive contact with it. According to representationalism, when we recall an event, we are aware of an imperfect mental idea of it that falls short of direct cognitive contact. Causal intermediaries between the experience and its recall may alter the content of the memory. The cognitive neuroscience of memory endorses some version of representationalism. Still, the idea we have of the experience is not just a mental process but also a neural process involving varying degrees of integrated information in the brain. The key causal intermediaries influencing the extent to which the representation resembles the experience are changing contextual factors in the time between the experience and the initial memory of it and later retrieval and reconsolidation of the memory.

Those who accept direct realism or representationalism tend not to fully appreciate the concept of memory as a dynamic process of continuous updating of information. They fail to appreciate the extent to which we consciously and unconsciously edit memory in our mental life. They focus too much on the "pastness" of memory and not enough on its future-oriented aspect. It is important to point out, though, that not all philosophers focus primarily on the past in discussing memory. Influenced by contemporary psychologists investigating memory, an increasing number of philosophers focus on the constructive aspect of memory and its neural correlates (e.g., De Brigard, 2013, 2017; Michaelian, 2016, pp. 82–85).

I do not engage with the literature on the politics of collective memory either (Blustein, 2008; Campbell, 2014; Stone and Bietti, 2016; Belavusau and Gliszczynska-Grabias, 2018). While memory of historical injustice toward certain groups generates an obligation for societies to restore justice, this type of memory is very different from the type I examine in this book. My concern is not with how groups remember the past but with how individuals remember it. I discuss memory as a neurobiological and psychological process rather than a social and political one. The general focus is on how information as a preserved neural and mental representation of the past shapes how we think and act in the present and future.

Normative questions about memory fall within the domain of neuroethics. This is an interdisciplinary field at the intersection of the clinical neurosciences, cognitive science, psychology, radiology (neuroimaging), philosophy and law. In a seminal paper published in 2002, Adina Roskies distinguished between two branches of neuroethics: the ethics of neuroscience and the neuroscience of ethics (Roskies, 2002). The first branch considers the potential benefits and risks to patients and research

subjects whose brains are mapped or monitored by structural and functional imaging. It also considers the potential benefits and risks of altering the brain with psychotropic drugs, surgery and electrical stimulation. The ethics of neuroscience also considers the obligations of clinicians and investigators to protect patients and research subjects with neuropsychiatric disorders from harm. The neuroscience of ethics generally pertains to the neurobiological basis of rational and moral decision-making. It concerns the cognitive and emotional capacity to consider reasons for and against brain interventions, how one may be affected by them and how one may make informed decisions to receive or refuse treatment and participate or decline to participate in research. While acknowledging that the ethics of neuroscience and the neuroscience of ethics "can be pursued independently to a large extent," Roskies noted that "perhaps most intriguing is to contemplate how progress in each will affect the other" (2002, p. 21).

An example of research on a disorder of memory capacity illustrates how the two branches of neuroethics can overlap. Suppose that a researcher conducting a clinical trial on deep brain stimulation of the hippocampal-entorhinal circuit as a potential therapy for anterograde amnesia recruits a patient with this disorder. The researcher is obligated to obtain informed consent from the research subject and ensure that he is not exposed to more than minimal risk from the intervention (ethics of neuroscience). Although memory systems often interact, they are dissociable. While the hippocampal damage impairs the subject's ability to form new memories, the prefrontal region mediating executive functions associated with the subject's working memory may be intact. Other neocortical regions storing episodic and semantic information used in working memory may be intact as well. This may provide the subject with enough cognitive and emotional capacity to consider the prudential and moral reasons for participating in research. He may be able to consciously hold information long enough and exercise a sufficient degree of decisional capacity to give informed consent to participate in the trial. Yet if he has a dysfunctional prefrontal cortex and his working memory is impaired, then he may lack the cognitive capacity to make decisions and give informed consent (neuroscience of ethics). This is one of the challenges in recruiting early-stage Alzheimer's patients for research on techniques designed to improve memory function. Changes in their brains affecting working memory may impair the degree of decisional capacity necessary to consent to participate as subjects in clinical trials.

A different example of research on a disorder of memory content further illustrates the overlap between the ethics of neuroscience and the neuroscience of ethics. Suppose that a researcher conducing a clinical

trial using a protein synthesis inhibitor to erase a pathological fear memory in the amygdala recruits a patient with PTSD as a research subject. Again, the researcher is obligated to protect the subject from harm by obtaining informed consent from her and ensuring that she is not exposed to more than minimal risk (ethics of neuroscience). The subject may have enough prefrontal-mediated cognitive function to process information about the trial, consider how she might contribute to and be affected by the trial and give informed consent to participate in it. However, if hyperactivation of her fear memory system projects to and impairs prefrontal function, then her cognitive capacity to weigh the potential benefits and risks of participating in the research could be impaired (neuroscience of ethics). Her hyperactive emotional state could compromise her capacity for informed consent. How one assesses questions in the neuroscience of ethics about whether a person affected by a memory disorder has the requisite decisional capacity may depend on the severity of the disorder. This capacity can be measured by combining neuroimaging or electrophysiological recording with assessment of the person's behavior.

The neuroethics of memory can be construed broadly to include not only questions about the potential benefit and harm of memory-modifying interventions in the brain. It also includes the effects of memory disorders and treatments for them on agency, identity and the role of memory in judgments of moral and criminal responsibility. Memory is thus relevant to issues in metaphysics, ethics and criminal law. These issues are informed by neurobiological and psychological determinations of memory function and dysfunction. An interdisciplinary perspective on memory corresponds to the interdisciplinary nature of neuroethics.

My discussion of the ethical and legal dimensions of memory is not driven by a single theory. Instead of selecting a theory and then discussing issues around it, I first raise the issues and then explicitly or implicitly apply a theory to explain why a course of action or policy is justified or unjustified. For example, whether memory should be enhanced or erased drives the application of an ethical theory to answer these questions. Because improvement in the capacity to consolidate and retrieve memories is the intended outcome of drugs or neurostimulation for disorders of memory capacity, the relevant ethical theory to assess these interventions is consequentialism. This same theory can provide a normative framework for addressing questions about the justification of erasing memories in disorders of memory content. In criminal law, the question of whether a victim of a crime with a traumatic memory of it has the right to erase the memory or is obligated to retain it to testify against the

perpetrator involves balancing deontological and consequentialist considerations. One must balance the victim's right to eradicate the source of continuous harm against the public's interest in avoiding potential future harm from the offender. Both deontological and consequentialist considerations may also be relevant to questions of whether a person who committed a crime but cannot recall his action should be held responsible or punished for it.

Methodologically and structurally, the book is divided roughly into two parts. I first outline some of the history behind theories of memory and describe different memory systems and mechanisms of encoding, consolidating, storing, retrieving, reconsolidating and re-storing memories. These systems and mechanisms constitute the neurobiological and psychological framework of memory. Within this framework, I examine the role of normal and abnormal memory in human thought and behavior and some of the philosophical and legal issues that arise from measuring and modifying it. Many of these issues involve disorders of memory content and memory capacity. Both types of disorder involve dysfunction at earlier and later stages of the memory process. Although their effects can be different, they can be equally mentally disabling.

In Chapter 1, I trace the main historical developments in theories of memory from Aristotle's concept of recollection to the current model of episodic memory as a constructive and reconstructive process. This is the legacy of Frederic Bartlett's early twentieth-century concept of memory as a reconstruction rather than reproduction of the past. Then I describe the taxonomy of memory systems. I also describe memory as a process that extends from encoding and consolidation in earlier stages to retrieval and reconsolidation in later stages. I explain how different neural circuits regulate different stages in this process. Retrieval and reconsolidation are critical to reconstructing episodic memories because they enable us to update memories to make them relevant to our present and future circumstances. The interconnection between retrieval and reconsolidation in destabilizing and restabilizing memories is necessary to update them. In addition to making memory adaptive, retrieval is the stage of the memory process where altering the content of memories, or erasing them, may be possible because they are labile and susceptible to change at this stage.

I discuss the role of memory in agency and identity in Chapter 2. Working memory is necessary for executive functions in reasoning and decision-making. It does not operate alone but relies on stored episodic and semantic memory. Because agency also involves goal-directed behavior and future planning, prospective memory is also critical for agency in holding intentions over time. Impairment in any of these types

of memory can interfere with the capacity to form and translate intentions into actions. Agency illustrates how different memory systems interact to enable the cognitive and emotional functions necessary for flexible thought and behavior. In addition, I discuss the role of episodic memory in the psychological connectedness and continuity that constitute personal identity. I describe how anterograde and retrograde amnesia can disrupt these psychological relations and thus identity. We constantly change the content of our memories by updating them. If the purpose of memory is not to reproduce the past but to enable us to simulate events in adapting to current and future environments, then only a critical core set of representations of the past may be necessary to provide a basis for these mental acts.

I explore the implications of radical life extension for memory and what it would mean for identity and self-regarding concern about the future. Specifically, I consider whether a person with a substantially longer life would become a different person with different interests beyond a certain point because of changes in the content of her episodic memories. If the adaptive function of memory involves changing its content and weakening of psychological connectedness and continuity over time, then adaptability may come at the expense of personal identity in a radically extended lifespan. I also discuss the loss of memory in dementia and the concept of precedent autonomy. This involves the question of whether a competent patient's wishes about life-sustaining or life-ending interventions apply when she is demented and has no memory of these wishes. If one accepts precedent autonomy, then the moral and legal force of a request made by a competent person in an advance directive transfers from the earlier to the later time.

In Chapter 3, I examine empirical, epistemological and ethical issues surrounding anesthesia awareness with postoperative recall. Some patients unexpectedly become aware during surgery despite receiving general anesthesia. If awareness cannot be detected intraoperatively, then a report from the patient may be the only way to confirm awareness. Yet a report of the experience requires a memory of it, and not all patients recall being aware. This raises the question of whether a patient could be harmed by becoming aware if she did not remember the experience. In cases where awareness is detected at a very early stage, we need to ask whether an anesthetist would be justified in infusing an amnesic drug to prevent consolidation of a memory of being aware without the patient's prior consent.

I consider whether an anesthetist would be justified or obligated to preoperatively inform patients of the possibility of awareness and such a preventive intervention as part of the consent process. When a patient forms a memory and recalls being aware, there may be reasons for and

against taking a drug that hypothetically could erase it. These issues are complicated by studies indicating that anesthesia can have different effects on episodic and fear memories. They are also complicated by research showing that the longer a memory has been consolidated and stored in the brain, the more difficult it is to weaken or erase it. In addition, patients may form implicit memories that are immune to the amnesic effects of anesthesia even if they are not aware during surgery. Explicit and implicit memories formed during surgery may have long-term harmful effects on patients.

In Chapter 4, I discuss disorders of memory content and interventions to treat them. These include psychiatric disorders such as PTSD, anxiety, depression, phobia and panic. They may develop from anesthesia awareness with recall or from other factors. The emotionally charged content of a fear memory of a stressful or traumatic event persists in the brain and mind beyond any adaptive purpose. Behavioral techniques such as extinction training, and pharmacological interventions such as propranolol, have been used to weaken the emotional content of these memories or replace them with other memories. But these interventions would not rule out the possibility of reactivation of the emotional representation. Protein synthesis inhibitors infused into the basolateral amygdala may block reconsolidation and effectively erase these memories. More invasive deep brain stimulation or focused ultrasound of localized nuclei in this brain region might also erase them. Even if these interventions could erase a pathological fear memory, they would have to be selective enough not to affect normal semantic, episodic and emotional memories. Because this is still very much hypothetical, it is not known how selective memory erasure could be and whether both maladaptive and adaptive memories would be affected.

I explain how erasing some memories would not necessarily alter identity and spell out differences in the ethical justification of erasing traumatic versus unpleasant memories. Memory modification can be consistent with authenticity in a person who decides to undergo it. Just as the content of our memories is constantly being updated, so too our authentic selves are never complete but constantly being revised and reshaped by us in our interaction with the environment. Some types of memory modification may be compatible with and complement natural memory updating as part of its adaptive purpose. However, it is not known what the short- and long-term neurophysiological and psychological effects of tinkering with memory would be. This underscores the need for placebo-controlled studies to determine the feasibility, safety and efficacy of erasing memories and the circumstances in which it would or would not be justified.