Beyond the Reason-Emotion Divide:

Philosophical Theories of Autonomy from a Neuroscience Perspective

by

Darlene Drecun

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Philosophy

Waterloo, Ontario, Canada, 2026

© Darlene Drecun 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Myrto Mylopoulos Associate Professor, Carleton University
Supervisor	Patricia Marino Professor, University of Waterloo
Internal Member	Chris Eliasmith Professor, University of Waterloo
Other Member	Mathieu Doucet Associate Professor, University of Waterloo
Internal-external Member	Ori Friedman Professor, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This dissertation explores how recent neuroscience research might bear on philosophical theories of personal autonomy, with a special focus on the work of Christine Korsgaard and Harry Frankfurt. A central question within the personal autonomy literature is what gives our actions self-governing authority? Many approaches have answered this question with requirements about how we ought to reason or feel about our desires. One prominent way of dealing with this question of authority is a requirement of reason being "in control" of our emotions, presenting the relationship between reason and emotion as competitive processes fighting for control over our decisions. However, by examining recent neuroscience research related to the concept of autonomy, I argue that this research paints a different picture, where reason and emotion function cooperatively rather than antagonistically. Furthermore, the research suggests a prominent role for emotion in many different autonomy-related processes.

In order to widely capture the many processes that underlie autonomy, I discuss the neuroscience of decision-making, self-control, voluntary action, and the conscious feelings related to agency and ownership over our actions. I examine some of the emerging trends in cognitive neuroscience that suggests that complex behaviours like decision-making and self-control emerge from large-scale neural networks located across widely distributed areas of the brain, where cognitive, emotional, and motivational information are deeply integrated at the neural level. I argue that this new notion of cognitive, emotional and motivational information as integrated in various processes, such as decision-making and control, has a direct impact on the concept of personal autonomy. First, this integration suggests that cognitive, emotional and motivational information all cooperatively contribute to autonomy processes, rather than competing. Second, the research suggests a larger role for emotion and motivation in processes like decision-making and self-control than is commonly assumed in philosophical theories of autonomy. Therefore, I argue that this integrated neuroscientific perspective highlights some important tensions between the neuroscience research on autonomy processes and philosophical theories of personal autonomy, like those of Korsgaard and Frankfurt.

I examine the possible implications that the neuroscience of decision-making, self-control, voluntary action, and conscious feelings of agency may have on the autonomy theories of Korsgaard and Frankfurt. I point out several key tensions between the neuroscience research and how these theories of autonomy understand desires as motivating our actions, the role of emotion in decision-making and self-control, and whether we ought to rely on our conscious feelings of agency and control over our actions when determining whether our actions are autonomous. I suggest overall that an integrated neuroscientific understanding of the processes that support autonomy-related behaviours can provide a novel approach to understanding the concept of personal autonomy.

Acknowledgements

I'm truly grateful for all the support that I received throughout the process of writing this dissertation. I would first like to thank my supervisor, Patricia Marino, for your unwavering support and encouragement throughout this process. Your guidance and feedback throughout the many conversations about my work was always helpful and your support was instrumental in getting me to the finish line.

I would like to thank my internal committee members, Chris Eliasmith, Matt Doucet, and my external committee members, Ori Friedman and Myrto Mylopoulos. Your feedback on my work was extremely insightful and helped me refine key ideas in this dissertation. Your comments also provided many new ideas for me to explore in my future work.

I would also like to thank Paul Thagard. Your support and guidance when I first started looking into the role of emotion in the neuroscience of decision-making was essential to the development of my initial ideas about this dissertation topic. I'm also very thankful to you for introducing me to many of the authors I've used to help shape my arguments in this dissertation.

And finally, I would like to thank my family and friends for your consistent support and encouragement over the years. I wouldn't have been able to complete this without you. Thanks to my oldest friend Harinder Bisla for her constant check-ins, and encouragement. I would also like to thank my niece, Sierra Joseph, for the many helpful conversations and support over the years. And lastly, to my husband Zachary Fouchard. I can't thank you enough for the many philosophical conversations over the years, as well as the constant support and encouragement throughout this whole process.

I'm grateful to have had so many cheerleaders in my life, and my deepest, heartfelt thanks to everyone.

Dedication

Dedicated to my parents,
for their endless support and encouragement over the years.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Abstract iv

Acknowledgements v

Dedication vi

List of Abbreviations x

Introduction 1

Interactive and Integrated Neural Network Approaches to Brain Function 4

Chapter Overviews 13

Chapter 1 20

Reason and Emotion in Theories of Personal Autonomy 20

Introduction 20

1.1 Frankfurt's Structural Theory of Autonomy: A Humean Approach 21

1.2 Watson's Structural, Reasons-Responsive Theory of Autonomy: A Platonic Approach 36

1.3 Korsgaard's Reasons-Responsive Theory of Autonomy: A Kantian Approach 41

Conclusion 50

Chapter 2 53

The Neuroscience of Decision Making: Reward, Motivation, and Value 53

Introduction 53

2.1 Perceptual Decision-Making Versus Value-Based Decision-Making 54

2.1.1 Perceptual Decision-Making 54

2.1.2 Value-Based Decision-Making 58

2.1.3 Goal-Directed Behaviours and the Basal Ganglia 60

2.2 The Orbitofrontal Cortex, the Striatum, and Value-Based Decision-Making 64

2.2.1 The Orbitofrontal Cortex and Value Representations in Decision-Making 64

2.2.2 The Striatum and Dopamine Signaling in Motivation and Reward Processes 68

2.2.3 Motivational Value Driving Goal-Directed Behaviours: Differences Between "Wanting" and "Liking" 72

2.3 A Shared Neural Mechanism for Decision-Making: Integration of Cognitive, Emotional, and Motivational Information 75

Conclusion 77
Chapter 3 80
The Neuroscience of Control: Goals, Emotion, and Motivation in Control Processes 80
Introduction 80
3.1 The Neuroscience of Self-Control 81
3.1.1 Self-Regulation 82
3.1.2 Self-Control 92
3.1.3 Cognitive Control 100
3.2 Control as a Duality: Emotion in Self-Control Dilemmas 111
Conclusion 119
Chapter 4 123
The Neuroscience of Voluntary Actions and the Experience of Agency 123
Introduction 123
4.1 Libet's Experiments and the Neuroscience of Free Will 124
4.1.1 Brain Circuits for Voluntary Action 127
4.1.2 Brain Circuits for Intentional Inhibition 129
4.1.3 Has the Neuroscience of "Volition" Been Captured? 130
4.2 The Feeling of Control: Conscious Awareness of Agency and Ownership 136
4.2.1 "Autonomy Disturbances" and the Brain 136
4.2.2 The Conscious Experience of Intentions and Control over our Actions 141
Conclusion 142
Chapter 5 146
Korsgaard's Theory of Autonomy: Exploring the Implications of Integrating Cognition, Emotion, and Motivation 146
Introduction 146
5.1 Korsgaard's Reasons-Responsive Account of Autonomy and the Role of Emotions 147
5.2. Self-Constitution in Korsgaard's Theory of Autonomy 156
5.3. The Role of Desires and the Motivation Process 162
5.4 Reflective Distance: Are we Consciously Aware of our Desires? 167
Conclusion 171
Chapter 6 174

Frankfurt's Theory of Autonomy: Emotions, Desires, and the Experience of Agency 174
Introduction 174
6.1 Reason and Emotion in Frankfurt's Theory of Autonomy 175
6.2 Desires, Goals, and our Motivation System 180
6.2.1 Frankfurt's Hierarchy of Desires and the Reward System 181
6.2.2 Frankfurt's Caring Approach to Autonomy: Essential Volitional Character as the Source of Authority 186
6.3. The Unwilling Drug Addict and the Neuroscience of Addiction 191
6.4 The Subjective Experience of Agency and Control 194
6.4.1 Experiencing "Autonomy Disturbances" 197
6.4.2 The Experience of Agency and Control: Implications for Frankfurt 199
Conclusion 205
Conclusion 208
Future Directions 215
References 218
Glossary 239

List of Abbreviations

ACC	Anterior Cingulate Cortex
aMCC	Anterior Midcingulate Cortex
BG	Basal Ganglia
BLA	Basal Lateral Nucleus of Amygdala
CMS	Cortical Midline Structures
DA	Dopamine
dACC	Dorsal Anterior Cingulate Cortex
dMFC	Dorsal Frontomedian Cortex
dlPFC	Dorsal Lateral Prefrontal Cortex
DS	Dorsal Striatum
HA	Histamine
IFG	Inferior Frontal Gyrus
IFGpo	Inferior Frontal Gyrus Pars Opercularis
area LIP	Lateral Intraparietal Area
lPAG	Lateral Periaqueductal Gray
mOFC	Medial Orbitofrontal Cortex
mPFC	Medial Prefrontal Cortex
area MT	Middle Temporal Area
NA	Noradrenaline
NAc	Nucleus Accumbens
OFC	Orbitofrontal cortex
PAG	Periaqueductal Gray
PFC	Prefrontal Cortex
pre-SMA	Pre-Supplementary Motor Area
M1	Primary Motor Cortex

RP	Readiness Potential
rIFC	Right Inferior Frontal Cortex
RCZ	Rostral Cingulate Zone
5-HT	Serotonin
SMA	Supplementary Motor Area
SN	Substantia Nigra
SNc	Substantia Nigra Pars Compacta
STN	Subthalamic Nucleus
VP	Ventral Pallidum
vlPFC	Ventrolateral Prefrontal Cortex
vmPFC	Ventromedial Prefrontal Cortex
VS	Ventral Striatum
VTA	Ventral Tegmental Area

"Reason and spirit…will preside
over the appetitive part."

—Plato, Republic

"Reason and it alone can produce
a will that is *good in itself*."

*—Kant, Groundwork for
the Metaphysics of Morals*

"Reason is, and ought only to be
the slave of the passions..."

*—Hume, A Treatise of
Human Nature*

Introduction

Following in the long philosophical tradition of Plato and Kant, many have argued that a purely rational will must be the motivational source of an agent's autonomous actions. A notable exception to this tradition came from Hume, where he compellingly argued that only the passions can be the motivational source of our will. These fundamental ideas about agency and the will can be found throughout the philosophical autonomy literature, where we regularly encounter the idea that we must choose between reason or emotion as the source of our actions when exercising our autonomy. Furthermore, there is a related idea embedded within the autonomy literature that insists that a person's reason must be in control of their emotions as a requirement for autonomy. One central aim of this dissertation is to challenge these notions about reason and emotion within the autonomy literature by examining current neuroscience research which, on the contrary, suggests that reason, emotion, and motivation processes are quite integrated in autonomy processes.

Autonomy, or self-government, is a central value in most moral and political frameworks. The capacity to self-govern is seen as a requirement for moral responsibility, political legitimacy, as well as the key requirement for medical and legal competency. The concept of individual or personal autonomy, generally refers to a person's individual capacity to act independently on their reasons, desires, and motives—free from external constraints (Christman 2003). It is personal autonomy that I will be focusing on throughout this dissertation. Different theories of personal autonomy focus on various components of this capacity (e.g., are there sufficient reasons to act, are the agent's desires authentic, has the agent's will been subjected to manipulation or external force), resulting in diverse requirements for individual autonomy. As a detailed discussion of all the various theories would go beyond the scope of this dissertation, I

will focus on three general conceptions of autonomy that are rooted in Platonic, Humean, and Kantian moral frameworks. This approach will give the topic of personal autonomy a wide breadth while also highlighting the key questions I will be considering about the role of reason and emotion in personal autonomy.

In particular, I focus on the autonomy theories of Harry Frankfurt (Frankfurt 1971; 1998c; 1999b) and Christine Korsgaard (Korsgaard 1996; 2009). Frankfurt's hierarchical theory of autonomy, following in the tradition of Hume's sentimentalism, is a prominent example of a more affective, desire-based account of autonomy. Frankfurt's early approach to autonomy focuses on the motivational structure of an agent as a hierarchy of desires. In his later works, there is a shift away from his structural approach toward a theory more focused on the authenticity of an agent's values. As will be discussed in detail in later chapters, Frankfurt's later work focuses on the motivational constraints of what an agent cares about.

In contrast, Christine Korsgaard's reasons-responsive theory of autonomy follows in the tradition of Plato and Kant, heavily prioritizing reason as necessary for both personhood and autonomy. Reasons-responsive theories of autonomy generally require that the agent's desires and personal motivations are responsive to reasons for their actions. Following this type of approach, Korsgaard's theory of autonomy requires an autonomous agent to be able to consciously reflect and deliberate about their desires and motivations, as will be discussed in greater detail in later chapters. Both Frankfurt's and Korsgaard's approaches rely on particular definitions of reason, desire, motivation, and personhood—all of which profoundly impact the way they define personal autonomy.

The contrast between Frankfurt's and Korsgaard's theories of autonomy highlights some of the key questions I will be considering in this dissertation. Some of these questions relate to

the self-government over our desires—such as, how theories of personal autonomy determine agential authority over our actions and how we know if our desires are really our own? These theories also raise questions about whether we can "reason" or "feel" our way to determining if our desires and motivations are authentically our own. Finally, a central question that runs throughout this dissertation is how reason and emotion are viewed within the personal autonomy literature. What role do they play in legitimizing our claim of autonomy and personhood, and therefore the authority we have over our own actions?

Neuroscience can help shed some light on some of these questions as well. There has been extensive work on the neuroscience of free will[1] and agency[2] (specifically, the phenomenological "feeling of agency"). However, there is a gap in the literature when it comes to the neuroscience of personal autonomy, with little published on this topic at this point[3]. One aim of this dissertation is to begin to address this gap and consider what the empirical literature can contribute to our philosophical understanding of personal autonomy. I will use current neuroscience research to explore how we make decisions and how we control and act on our desires. I also consider the ways in which Korsgaard and Frankfurt view central concepts such as desire, motivation, emotion, and reason and whether these are compatible with current neuroscientific understandings of these concepts. Finally, I will consider possible tensions

[1] For the neuroscience of free will see: Libet, B., Gleason, C.A., Wright, E.W., and Pearl, D.K. "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act." (Libet et al. 1983); Wegner, D.M. *The Illusion of Conscious Will* (Wegner 2002); Haggard, P. "Human Volition: towards a neuroscience of will" (Haggard 2008).

[2] For the neuroscience of agency and the "feeling of agency" see: Haggard, P. "Sense of Agency in the Human Brain" (Haggard 2017).

[3] For the neuroscience of autonomy see: Felson, G., and P. B. Reiner "How the neuroscience of decision making informs our conception of autonomy" (Felsen and Reiner 2011); Peterson, A. "Should Neuroscience Inform Judgments of Decision-Making Capacity?" (Peterson 2019); Levy, N. "Addiction, Autonomy, and Informed Consent: On and Off the Garden Path" (Levy 2015).

between the empirical evidence and the central claims proposed by Korsgaard and Frankfurt. There may be a way to reconcile aspects of Korsgaard's and Frankfurt's approaches with the neuroscience literature. However, the focus of this project is to point out areas where these tensions between the empirical research and the philosophical theories occur and to consider how the neuroscience literature challenges traditional ideas within the philosophical autonomy literature.

Interactive and Integrated Neural Network Approaches to Brain Function

An analysis of the capacity for personal autonomy from a neuroscientific perspective necessarily involves looking at many different processes across widely distributed areas of the brain. There is not a specific "autonomy" process or region of the brain, so it is necessary to look at the different processes that make up what we think of as autonomous behaviour. I chose to focus on decision-making and self-control as two central behaviours that capture many elements of what it means to act autonomously. Much of the philosophical literature discussing personal autonomy questions how we make autonomous decisions, but also, how we ought to deal with self-control dilemmas. Self-control dilemmas question the authenticity of our desires when there is a desire conflict—asking how do we know what we *really* want? Self-control dilemmas often present a conflict between desires where one desire is overwhelmingly tempting, while the other desire is more strongly linked with a long-term, reflective desire. These scenarios are meant to show how we overcome tempting desires and pursue the desire we truly want, signifying an autonomous choice through the agent's effective self-control or "will power". Therefore, I focus on both decision-making and self-control as two central autonomy-related processes.

Typically, when the capacity for autonomy has been discussed from a more empirical perspective (e.g., from within the bioethics and neuroethics literature), the few articles published so far have primarily focused on the neuroscience of decision-making alone. I think both decision-making and self-control processes should be considered to cover more widely the elements of the philosophical personal autonomy literature and is therefore, a good place to begin when considering the underlying neuroscience of personal autonomy. However, from a neuroscientific perspective, there are many different components at work that support our decision-making and self-control processes.

As a starting point for considering the many different components of autonomy, I chose to focus on the neural level rather than other possible levels of analysis (e.g., computational or psychological levels). I was interested in investigating the growing empirical research on the neuroscience of decision-making and control processes in order to see how they would relate to the philosophical concept of autonomy. I argue that looking at brain function from the neural perspective can provide new information about the mechanisms that underlie autonomous decision-making and control. Additionally, the neural level highlights some key components about brain function that are not as apparent at other levels of analysis. For example, the fact that recent research has found that emotion processing is integrated into what was traditionally considered cognitive or motivational processes, such as memory or reward circuits (Beam et al. 2021), cannot be seen at other levels (e.g., such as the psychological level, for instance). I argue that this integration occurring at the neural level has implications for the way we understand autonomy processes, as well as the potential to move us in a new direction in how we understand the concept of autonomy.

Given the complexity of brain function at the neural level, it is important to look at the function of the brain as a whole and dynamic system, how each of these processes interact and integrate with each other, and finally, how this results in autonomous behaviours emerging from these processes. Using a "whole brain" approach is adopted here in contrast to "functional localization," an approach that aims to match one area of the brain (or even one network) with a psychological state or behaviour. Throughout this dissertation, I will be focusing on frameworks that consider the brain as a whole instead of approaches that are rooted in functional localization.

Historically, brain function was investigated from a very localized perspective (see Pessoa 2013; Pessoa, Medina, and Desfilis 2021 for a review of this idea and a history of functional localization). A functional localization approach maintained that certain areas of the brain were thought to be the source of specific functions or behaviours (e.g., "fear" resides in the amygdala, or "reason" resides in the prefrontal cortex). This type of one-to-one mapping of the brain (one brain region to one function) led to a highly modular understanding of the neuroarchitecture of the brain (For a critique of one-to-one mapping see Anderson 2014; Poldrack 2018; Pessoa 2022).

However, we should be wary when analyzing neuroimaging studies of this type of one-to-one mapping, as it can sometimes lead to what has been called the "reverse inference fallacy" (Poldrack 2006; 2018). The problem Poldrack points out is that a neuroimaging study may show activation of a specific brain area when the participant is engaging in some sort of behaviour or psychological state. However, we cannot infer based on this alone, that the area that is showing activation is the source of that behaviour or psychological state. For example, the anterior cingulate cortex (ACC) is known to be involved in conflict resolution tasks. However, if the ACC is activated in a neuroimaging study, we would not be able to infer from that activation

alone that the participant is currently engaging in conflict resolution because activation of the ACC is present in many different cognitive functions (Poldrack 2018, 22). Therefore, this type of one-to-one mapping of a specific brain area and psychological state or behaviour can be problematic (Poldrack 2018, 22). This is a trap we do not want to fall into as well when trying to pull out information about brain function in relation to personal autonomy.

Historically though, this type of one-to-one mapping of the brain was a common approach. Early on, many efforts were made to try to localize the "emotional brain" and the "cognitive brain." Much of this early research focused on the prefrontal cortex (historically the prefrontal cortex was often termed the "neocortex") as the home of "cognition," and the "limbic system" (e.g., brain regions such as the basal ganglia, amygdala, hippocampus, thalamus, and hypothalamus) as the source of emotion in the brain[4]. These early approaches viewed brain architecture in not only a highly modular way, but also a strictly hierarchical way—where the higher "neocortex," or the seat of reason, was understood to be in control of the lower, more primitive, emotional areas of the brain.

The impact that this specific type of hierarchical theory has had on the way we think of personhood and the mind, where our "higher reason" controls our "lower emotions," has been monumental and is so deeply embedded in our historical understanding of human life. We can see similar hierarchical ideas present in early conceptions of the hierarchical soul espoused by Plato (Plato 1997) (e.g., *The Republic* and *Phaedrus*) and Aristotle (Aristotle 2009) (*De Anima II*) to later scientific models of the hierarchical mind by scientists such as Charles Darwin (1872) (Darwin 2014), John Hughlings Jackson (Jackson 1884), Ludwig Edinger (Edinger and Rand

[4] The term "limbic" was coined by Paul Broca in 1878 originally. It was also famously used by James Papez as far back as 1937 when describing the role of emotion and the "Papez Circuit". But the term 'limbic system' is usually attributed to Paul D. MacLean (1949, 1952), who reconceptualized the Papez Circuit and coined the term "limbic system."

1908), James Papez (Papez 1937), and many more. From these early scientists, the application of a hierarchical organization on the brain were primarily evolution-based, where it was generally thought that new parts of the human brain slowly evolved more and more complex brain structures on top of the older primitive ones, shared by mammals and reptiles.

While each specific evolutionary theory was unique in their own way, the running theme underlying many of these approaches was the idea that only humans evolved into animals with a neocortex, and therefore are uniquely capable of rational, higher cognition. However, there is also a precariousness to this rationality, where one must have control over the lower, primitive parts, like our emotions and instincts (a.k.a., our "inner beast"). Without this type of control over our emotions and instincts, we are no better than the reptiles. We can see a direct line of reasoning, from early philosophical accounts like Plato's tripartite soul, and these 19th and 20th century evolutionary accounts of the brain—where "reason," by way of the neocortex, is in control of our "primitive" emotions and impulses. Relatedly, we can see similar ideas within early faculty psychology[5] approaches to the mind, where it was thought that the mind could be carved up into different "faculties" or "modules," each with their own separate mental function[6] (e.g., emotion, cognition, perception, consciousness).

These early modular and hierarchical ideas about the mind led to highly influential psychological theories in the 20th century, like the "triune brain," coined by MacLean in the 1960s (MacLean 1990). The triune brain views the human brain as governed by the more

[5] Faculty psychology approaches are attributed to early ideas by philosopher Thomas Reid (1710-1796), phrenologist Franz Joseph Gall (1758-1828), and also revived as "massive modularity" in the 1980s by Jerry Fodor (Fodor 1983) and by Leda Cosmides and John Tooby (Cosmides and Tooby 1992). Faculty psychology is still a popular approach to the study of emotions in psychology today (See Lindquist and Barrett 2012 for a discussion and critique of the faculty psychology approach).

[6] See (Anderson 2014) and (Lindquist and Barrett 2012) for a detailed discussion of this approach.

"evolved" neo-mammalian cortex (where cognition and complex reasoning takes place), over and above the paleo-mammalian limbic system (which generates emotions) and the reptilian brainstem (which generates instincts and impulses) (MacLean 1990). The concept of the triune brain again can be directly traced back to the hierarchical ideas of Plato's tripartite soul and Darwin's evolutionary theories of the human brain, as applied directly and hierarchically onto the brain (Barrett and Satpute 2019, 2n4).

Theories like the triune brain still have a strong hold on popular thinking about the brain. It is commonplace to hear references to our "reptile brains" as code for our impulses in everyday life. The idea of the triune brain can also still be seen to play a prominent role in some current psychological theories. Additionally, the influence of these localized function approaches to the mind can also still be seen in current neuroscientific research. As discussed recently by Pessoa, Medina, and Desfilis (2021),

> More than a century later, knowledge about the brains of vertebrates has expanded enormously. Yet, old thinking dies hard. Antiquated views of brain evolution continue to influence neuroscience, even if implicitly. As an example, consider that most frameworks of brain organization are heavily centered on the cortex. These descriptions view 'newer' cortex as controlling subcortical regions, which are assumed to be (relatively) unchanged throughout evolution. (Pessoa et al. 2021, 3)

More recent work in comparative anatomy looking at the evolutionary changes in brain architecture have demonstrated a reorganization of existing circuits in the brain and expansion of certain key brain regions, such as the basal ganglia (Pessoa, Medina, and Desfilis 2021). This new conception has replaced the historical additive view of the brain—the view that the evolutionarily new cortex was added on to the older, unchanged parts of the brain such as the brainstem, or amygdala, for instance (see Pessoa, Medina, and Desfilis 2021 for a complete overview of this research). However, as Pessoa, Medina, and Desfilis mention in the above

quote, “old thinking dies hard” (Pessoa et al. 2021, 3). We can still see the influence of these early accounts within the neuroscience literature on decision-making and self-control, and more widely in discussions of emotion, cognition, and motivation—which I will argue directly impacts the way we think about personal autonomy.

Moving away from these early ideas, new and innovative work on the architecture of the brain have suggested that neural networks can be understood as widely distributed in the brain, dynamically interacting to produce our complex behaviours (Lindquist et al. 2012; Anderson 2014; Barrett and Satpute 2019; Hoemann and Barrett 2019; Pessoa 2022). It has also been suggested that cognition, emotion, and motivation may be highly integrated in order to help support many of our complex behaviours (Pessoa 2022). What exactly does an interactive and integrated neural network approach mean in terms of brain connectivity? This involves looking at the architecture of the brain both in terms of its structural and functional connectivity. According to some prominent work on connectivity and brain architecture, we can understand the brain as being divided into a set of discrete elements (nodes) and their mutual connections (edges) that jointly form an object called a graph (Sporns 2011). These are often referred to as modules in a structural sense and focus on the anatomical connections within a network. The brain’s functional characteristics are thought to emerge from the dynamic interactions between the modules, where the brain continually and dynamically organizes and reorganizes itself into different functional patterns in order to perform certain tasks (Anderson 2014, 42). Although these views refer to modules, this term is used in a different sense than how it was previously used, such as with massive modularity approaches where a module represents one mental state or behaviour. As discussed recently by Anderson (2014),

> The evidence does seem to support the move away from massive modularity approaches, however, there are approaches that support a slightly different

> understanding of modularity. That is, there have been many analyses of the structural and functional connectivity in the brain that support the position that the brain is composed of many interacting modules. (Anderson 2014, 42)

As Anderson (2014) further explains, these "modules have well-defined abstract structural characteristics but flexible functional characteristics" (Anderson 2014, 42), with brain function then emerging from the dynamic interactions between these modules (Anderson 2014, 42). Although the question of modularity is still debated within the literature (see Shine and Poldrack 2018 for a detailed discussion), there is significant empirical support for the idea that anatomical network modules are involved in a wide variety of tasks (Cole et al. 2013; Anderson 2014; Price and Friston 2005). From this perspective, we can understand the brain as consisting in broadly distributed functional networks that interact to produce a wide-range of mental states and behaviours (Lindquist and Barrett 2012; Anderson 2014; Lindquist et al. 2012; Barrett and Satpute 2019; Hoemann and Feldman Barrett 2019). This approach views the construction of mental states as emerging from basic, domain-general networks (Lindquist and Barrett 2012). The networks are thought to dynamically reorganize into different functional patterns to perform a variety of different tasks—meaning that one mental state or behaviour does not directly map onto one particular network or brain region.

Why does the issue of modularity and the degree to which networks are engaged in a variety of tasks matter to autonomy? I argue that these interactive and integrated neural frameworks raise some new and interesting questions about the way we understand certain concepts and processes related to autonomy. For instance, approaches like those of Lindquist and Barrett (2012), Lindquist et al. (2012), or Pessoa (2022) view processes like cognition, emotion, and motivation as integrated and bound together in various, complex, and dynamic ways, rather than separate mental categories. This research suggests that emotional and motivational

processes are crucial in many cognitive tasks, and vice versa, seemingly breaking down the boundaries between these mental concepts (e.g., concepts like cognition, emotion, and motivation, and perception). Given that these are central components to processes like decision-making and self-control, they have a direct impact on how we understand the concept of personal autonomy.

Consequently, I argue that decision-making and self-control are best viewed as complex processes involving the integration of cognitive, emotional, and motivational information. The integration of emotional and motivational information into processes outside of traditional "emotion" or "motivation" domains contributes to the blurring of the traditional categories of "cognition," "emotion," and "motivation." This also moves us away from thinking about "reason" as competing with "emotion." If we are beginning to question whether these categories should be considered separate mental categories, this impacts the way we define their role in processes related to autonomy. I will explore the implications of this perspective on the way we understand autonomous decision-making and self-control, and therefore the way this might impact our understanding of personal autonomy. I will argue that the neural integration of cognitive, emotional, and motivational information described can lead to a new approach to personal autonomy—one where autonomous decision-making and self-control are driven by a more affective, rewards-based process.

In sum, throughout this dissertation I discuss the recent move in neuroscience toward a more interactive and integrated understanding of brain function and how this view impacts the way we understand personal autonomy. I also discuss the move away from one-to-one mapping of brain function, towards a many-to-many linking of multiple networks and circuits across various tasks and behaviours. This approach views brain function resulting from widely

distributed functional networks that interact, and from which our complex behaviours like decision-making and self-control emerge. This approach leads itself to a new understanding of decision-making and self-control, as well as concepts fundamental to these processes, such as emotion, cognition, and motivation. This, in turn, provides a new way of thinking of personal autonomy. The neural integration of cognitive, emotional, and motivational information leads to my later arguments that emotional and motivational value information are the main drivers of decision-making and control—not reason. However, this is not to say that the categories of emotion and motivation rule over reason. Rather, I argue that the integration of cognitive, emotional, and motivational information in functional networks moves us away from considering reason, emotion, and motivation as distinct mental categories. Instead, such integrated neural networks are what produce tasks like decision-making and self-control. As we will see, the neuroscientific research on decision-making and self-control suggests that a more affective, reward-driven valuation system underlies autonomy processes. As such, I will show how this type of neural integration perspective brings about a novel approach to the concept of personal autonomy.

Chapter Overviews

Chapter One provides an overview of three contemporary theories of personal autonomy: the theories of Harry Frankfurt (Frankfurt 1971; 1998c; 1999b), Gary Watson (Watson 1975; 2004), and Christine Korsgaard (Korsgaard 1996; 2009). While this dissertation focuses primarily on Frankfurt and Korsgaard, I chose to also include Watson's theory of autonomy in this chapter as an illustrative example of some of the key ideas discussed in Frankfurt's and Korsgaard's theories of autonomy. The three contemporary accounts broadly follow in the

historical traditions of Hume, Plato, and Kant, respectively. Hume, Plato, and Kant are known to have their own distinctive takes on how they conceive of autonomy and what it means to self-legislate one's actions. Their positions on this question relate to the way in which reason and emotion are portrayed in autonomy: that is, does having the authority to self-legislate one's actions require reason to be in control of our emotions? Or do emotions play a key role in motivating our autonomous actions?

In this chapter, I consider each of the three contemporary approaches and how their conceptions of personal autonomy are impacted by the way they define a role for both reason and emotion in the autonomy process. Each of these accounts provides a different answer to how we are supposed to determine the agential authority required of us for personal autonomy. By highlighting the historical traditions they stem from, I also show how their commitments to the way these traditions understand reason and emotion have had a large impact on the way their theories of autonomy developed.

The next two chapters turn to the contemporary empirical research on decision-making and self-control, as two main processes that are involved in what we think of as personal autonomy—making decisions for ourselves and how we control those decisions in our goal-directed behaviour. Within these chapters, I also consider how cognition and emotion have been historically understood in the neuroscience literature as well as the recent developments that have moved toward understanding many processes from an integrated cognitive-emotional-motivational perspective. Chapter Two reviews the neuroscience of decision-making, focusing on how the reward system helps compute value, thereby supporting our decision-making process. Importantly, in this chapter I discuss the recent neuroscience research that argues that we should understand decision-making as a highly integrated decision-making system, encompassing

widely distributed areas of the brain (cortical and subcortical) which involves cognition, emotion, and motivation processes.

Chapter Three provides a review of the neuroscience research on self-control. First, this chapter aims to provide a clear definition of three similar, but often conflated control processes: self-regulation, self-control, and cognitive control. I also review recent research that shows how goals, conflict, emotion, and motivation impact each of the three types of control discussed. Within this chapter, I also consider the idea of control as a conflicting duality between emotional and rational control. In particular, self-control dilemmas, which are prevalent in the philosophical autonomy literature, are often described as a conflict between deliberative, long-term goals competing with highly tempting, short-term rewards. I argue that the definition of self-control should be expanded beyond this narrow duality of long-term goals versus short-term rewards, and review evidence that supports this position. Additionally, I review some recent research that presents self-control as an emotional process. This new research raises questions about the way in which we have traditionally considered self-control as a highly deliberative, reason-based process.

Chapter Four examines the research focused on the neuroscience of volition. Specifically, I review some of the research on self-generated voluntary action and our conscious feelings of agency and ownership over our actions. Although this research is focused on the neuroscience of free will rather than autonomy specifically, I attempt to pull out some relevant information related to autonomy. Looking at some of the prominent experiments on free will, such as the experiments by Benjamin Libet and colleagues in the 1980s, I discuss whether this experimental paradigm tells us something about our everyday autonomous actions. This chapter also considers the related research on our conscious experiences of agency and ownership over our actions.

This research relates to the way we experience our autonomy, control, and ownership over our actions, and how we understand our identities as being held constant over time, amongst various environmental factors.

Overall, to critically examine the concept of autonomy from a neuroscientific perspective, I explore recent research on autonomy-related processes like decision-making and self-control, which I argue can help clarify our conversations about personal autonomy. The neuroscience chapters present new ways of thinking about the interactions between reason, emotion, and motivation and their roles in decision making, self-control, and our conscious feelings of agency. The empirical research helps develop and support my argument that we need to rethink these concepts in terms of their relationship with personal autonomy. I argue that the research indicates that the integration between cognition, emotion, and motivation suggests a mutually cooperative relationship that helps support behaviours like decision-making and self-control. I further maintain that the neuroscience research also strongly suggests a central role for affect in decision-making and self-control. I suggest that a more accurate representation of the way in which emotion and motivation affect decision-making and control processes needs to be translated into the autonomy literature. Therefore, I argue that an integrated neuroscientific approach highlights some important tensions between the neuroscience research on autonomy processes and the philosophical theories of personal autonomy that I explore in detail in the last two chapters.

Chapters Five and Six return to Korsgaard's and Frankfurt's theories of autonomy. This time I examine their theories in light of the reviewed neuroscience research in the preceding chapters. Chapter Five compares Korsgaard's reasons-responsive account of autonomy with the integrated (cognitive, emotional, and motivational) accounts of decision-making and self-control.

I point out some key areas where Korsgaard's account is not compatible with the neuroscience research on decision-making and self-control processes.

In particular, there are three key areas where there may be some tension between the empirical literature and Korsgaard's account of autonomy. First is the way in which Korsgaard defines emotions and how she perceives the role of emotion within her theory of autonomy. Second, Korsgaard's conception of personhood relies on a Platonic idea of self-control—with a problematic requirement of reason being in charge of our emotions. I explore both of these areas where Korsgaard discusses emotions and whether this is compatible with the neuroscience of emotions. Third, I consider how Korsgaard conceives of the motivational aspects of action in her theory, including her concept of "reflective distance." I compare this to the neuroscience research previously reviewed and argue that there are some key differences. Overall, I argue that there are some central assumptions that Korsgaard makes about emotional and motivational processes that may be fundamentally incompatible with what the neuroscience suggests.

Lastly, Chapter Six looks at Frankfurt's theory of personal autonomy in light of the neuroscientific research discussed in earlier chapters. Structural theories of personal autonomy, like Frankfurt's, focus on how desires rather than reason, motivate our autonomous actions. On the surface, this focus on desires motivating our actions seems to be more in line with the empirical research which shows an important role for cognitive-emotional-motivational processes in how we make decisions, motivate ourselves to act, and control our actions. However, I argue in this chapter that there are some important differences between the neuroscience of decision-making, control, our conscious feelings of agency, and Frankfurt's approach to autonomy.

First, I discuss the role of emotions in Frankfurt's account of autonomy and compare this with the integrated, cognitive-emotional-motivational frameworks for decision-making. I argue that the neuroscience research on decision-making does not support Frankfurt's conception of autonomous decision-making due to his dismissal of emotions as an authoritative motivational source. I also discuss Frankfurt's notion of desires and how they motivate and provide authority over our actions. Here, I point out some implications for Frankfurt's view based on the neuroscience research on the reward network. I follow this with a discussion of the neuroscience of addiction, comparing the empirical research to Frankfurt's well-known example of the unwilling drug addict. I argue that there are some important differences in the way addiction is understood that may impact a person with substance abuse disorder and their capacity for autonomy. Lastly, I discuss some of the main tensions between Frankfurt's approach and the neuroscience research related to the conscious feelings of agency and ownership. I consider in this section whether Frankfurt's reliance on "feeling satisfied" with our desires may present a problem for his theory of autonomy.

In conclusion, through an examination of current neuroscientific research on the processes of decision-making, self-control, voluntary action, and the conscious feelings of agency, I hope to point out some interesting tensions between the empirical research and prominent approaches to personal autonomy. I also consider the ways in which Korsgaard and Frankfurt view central concepts such as desire, motivation, emotion, and reason and whether these are compatible with current neuroscientific understandings of these concepts. As I'll explore throughout this dissertation, emerging trends in cognitive neuroscience suggest that complex behaviours emerge from large-scale networks widely distributed across the brain, from within which cognition, emotion, and motivation processes are often integrated in order to

perform tasks. I argue that this integrated neural perspective can provide a novel approach to understanding the concept of personal autonomy.

One key point I will explore further is whether the role of emotion has been under-played in both decision-making and self-control within the philosophical literature, therefore minimizing its role in a concept like autonomy. Going back to the initial questions I opened with, what is the role of reason and emotion in autonomy? Do reason and emotion need to compete? Does reason need to be in control of our emotions for us to be an autonomous agent? Or maybe neuroscience can show us that reason and emotion are, in fact, deeply and fundamentally integrated—consequently, making both reason and emotion fundamental components of personal autonomy in ways that have been underexplored.

Chapter 1

Reason and Emotion in Theories of Personal Autonomy

Introduction

Within the autonomy literature, there has been a long-held tradition of prioritizing the role of reason in moral action—from early Platonic accounts connecting reason to "the good,"[1] to the later Kantian ideas of a purely rational will as the source of autonomous action (Kant 1998). The one break in this long tradition was Hume's moral sentimentalism, where he argued that reason cannot motivate our actions. Rather, Hume's view was that only the passions can motivate our will (Hume 1985). These key ideas about reason and emotion, and in particular the battle between reason and emotion, have long influenced our thinking about autonomy.

The aim of this chapter is to provide an overview of several influential, contemporary accounts of autonomy. However, to do so, we must also recognize the underlying foundations of each of these approaches and the historical traditions from which they stem. For this reason, this chapter is divided into three main sections: Platonic accounts of autonomy, Humean accounts of autonomy, and Kantian accounts of autonomy. This chapter will review the autonomy theories of Harry Frankfurt, Gary Watson, and Christine Korsgaard, as prominent contemporary accounts that broadly follow in the traditions of Hume, Plato, and Kant, respectively. I will also examine the ways in which each of these theories of autonomy consider reason and emotion and how they might conflict. I suggest that within each of these contemporary approaches to autonomy, there are still remnants of the Platonic and Kantian idea that emotions play a negative role in

[1] Plato discusses the idea of the tripartite soul, where reason rules over spirit and appetite, in many different works; however, there are lengthy discussions of these ideas in *The Republic* and *Phaedrus.*

autonomous decision-making and action by acting as distractions from reasoned judgments. This chapter provides an overview of each of these accounts of autonomy, and how these approaches understand the role that is played by reason and emotion in the autonomy process.

1.1 Frankfurt's Structural Theory of Autonomy: A Humean Approach

A contemporary Humean approach to autonomy comes from Harry Frankfurt's early work on free will and moral responsibility. In "Freedom of the Will and the Concept of a Person" (Frankfurt 1971), Frankfurt outlines a conception of the self that characterizes an agent's structure of the will in the form of a split-level hierarchy of higher-and lower-level desires. According to Frankfurt, humans and animals share the capacity to form desires of the first-order: "desires to do or not do one thing or another" (Frankfurt 1971, 7). However, the capacity that is unique to humans is the ability to form second-order desires, where we can want or choose to desire something. On this account, autonomy requires an agent's second-order desires to identify with their first-order desires. According to Frankfurt, "someone has a desire of the second order either when he wants simply to have a certain desire or when he wants a certain desire to be his will" (Frankfurt 1971, 10). It is the desires that an agent wants to be his will, that Frankfurt calls "second-order volitions" (Frankfurt 1971, 10).

Furthermore, it is the second-order volitions that, for Frankfurt, are essential for personhood. To explain this idea, Frankfurt introduces the examples of the "wanton" and the "unwilling drug addict" (Frankfurt 1971). He uses this example to contrast the wanton's lack of personhood with the unwilling addict's capacity for personhood. The unwilling addict hates his addiction, but his desire to take drugs are too powerful and he eventually gives in to the desire. According to Frankfurt the unwilling addict has conflicting first-order desires of both wanting

and not wanting to take the drug. Although the unwilling addict has conflicting first-order desires, the fact that the unwilling drug addict hates his addiction indicates a higher-order desire to not take the drug. It is this higher-order desire that the unwilling drug addict wants to constitute his will (Frankfurt 1971, 12). However, even if the unwilling drug addict is unable to will this higher-order desire into an action because of his intense cravings, he is still granted personhood due to the very fact that he is capable of forming such higher-order desires (Frankfurt 1971, 12).

Frankfurt contrasts the unwilling drug addict with the wanton. The wanton drug addict does not care which of his first-order desires he acts on, and in fact, it doesn't even occur to him to consider which of his desires should constitute his will (Frankfurt 1971, 12). His lack of higher-order desires and volitions signifies the fact that he cannot be considered an agent as he has no identity beyond his first-order desires. As Frankfurt states, "[w]hen a person acts, the desire by which he is moved is either the will he wants or a will he wants to be without. When a wanton acts, it is neither" (Frankfurt 1971, 14).

For Frankfurt, the wanton's actions reflect the first-order desire to take the drug without being particularly concerned about the desires that move him to act. It could also be the case that the wanton has conflicting first-order desires about taking the drug. However, the wanton does not prefer that one desire over the other should constitute his will, unlike the unwilling addict (Frankfurt 1971, 14).

Through the formation of a second-order volition, the unwilling drug addict identifies with one of the conflicting first-order desires. By doing this, Frankfurt claims he makes the chosen desire more truly his own. The wanton, on the other hand, does not care which competing desire wins. Thus, this is a key distinction for Frankfurt—the unwilling drug addict has the

criteria for personhood, though would not be considered a person with free will if they gave in to the desire to do drugs, while the wanton does not achieve personhood. A free person is free to "have the will he wants" or is "free to will what he wants to will" (Frankfurt 1971, 14). A wanton does not will any desires for himself and therefore does not have personhood. Therefore, of central concern to Frankfurt's early work on autonomy, is determining whether a person's motivational structure is capable of second-order desires and volitions. If not, then they do not have the capacity to be autonomous.

Frankfurt's distinction of higher- and lower-order desires has been both extremely influential within the autonomy literature as well as highly criticized. One important challenge raised early on by both Irving Thalberg (Thalberg 1978) and Gary Watson (Watson 1975) has been called the infinite regress problem. This challenge raises questions about Frankfurt's prioritization of second-order desires. Thalberg primarily questions why we should we stop at the second level? Why can't there be third-level desires and volitions, and so on? Furthermore, why should we even consider the second-level more genuinely authentic than an agent's first-order desires (Thalberg 1978, 130)? As Thalberg argued, "why not go on to third-story or higher desires and volitions? And if that is somehow impossible, why grant that a second-order attitude must always be more genuinely his… than those you run into at ground level?" (Thalberg 1978, 130).

Watson also raises similar questions, particularly why one should necessarily care about one's higher-order volitions? He states that, "since second-order volitions are themselves simply desires, to add them to the context of conflicts is just to increase the number of contenders; it is not to give a special place to any of those in contention" (Watson 1975, 94). Frankfurt anticipates this issue and tries to provide a non-arbitrary reason to cut off the regress at the second-order

level. He argues that "when a person identifies himself decisively with one of his first-order desires, this commitment 'resounds' throughout the potentially endless array of higher orders"(Frankfurt 1971, 16).

In later work, such as "Identification and Wholeheartedness[2]" (Frankfurt 1998a) and "The Faintest Passion[3]" (Frankfurt 1999c), Frankfurt continues to address the infinite regress problem. Moving away from his initial approach focusing on the identification and endorsement of a higher-order desire, Frankfurt now argues that one of the conflicting lower-order desires must be decided on "wholeheartedly." To further explain what he means by wholehearted, Frankfurt brings up the problem of ambivalence as an example of not achieving wholeheartedness. For Frankfurt, a person who is ambivalent does not know what they want, and this is not because of some sort of cognitive problem or lack of information available to the person. Rather, Frankfurt states the ambivalent person's will is unformed (Frankfurt 1999c, 99). He goes on to say that "if ambivalence is a disease of the will, the health of the will is to be unified and in this sense wholehearted" (Frankfurt 1999c, 99). Thus, for Frankfurt, the idea of wholeheartedness indicates a unity of the will that signifies what the agent wants to be his will (Frankfurt 1999c, 99). Importantly, Frankfurt is clear that wholeheartedness is not a state of enthusiasm nor is it some sort of resoluteness of a person's volitional state (Frankfurt 1999c, 99). Wholeheartedness has to do with the organization of the will, not, as Frankfurt states, "its temperature" (Frankfurt 1999c, 99). In terms of how wholeheartedness applies to freedom of the will, Frankfurt states the following:

---

[2] "Identification and Wholeheartedness" originally published in 1987, reprinted in Frankfurt's collection of essays: Frankfurt, H.G. 1998. *The Importance of What we Care About*. Cambridge, UK: Cambridge University Press.

[3] "The Faintest Passion", originally published 1992, reprinted in Frankfurt's collection of essays: Frankfurt, H. G. 1999. *Necessity, Volition, and Love*. Cambridge, UK: Cambridge University Press.

> The dilemma can be avoided if we construe the freedom of someone's will as requiring, not that he originate or control what he wills, but that he be wholehearted in it. If there is no division within a person's will, it follows that the will he has is the will he wants. His wholeheartedness means exactly that there is in him no endogenous desire to be volitionally different than he is. Although he may be unable to create in himself a will other than the one he has, his will is free at least in the sense that he himself does not oppose or impede it. (Frankfurt 1999c, 100)

How does this new concept of wholeheartedness help Frankfurt avoid the infinite regress problem? For Frankfurt, wholeheartedness is tied to "satisfaction." He defines satisfaction as a "state of the entire psychic system—a state constituted just by the absence of any tendency or inclination to alter its condition" (Frankfurt 1999c, 104). In this sense, being satisfied does not require a person to have any particular beliefs, feelings, attitudes, or intentions; nor does is require any sort of approval or endorsement of actions. It is simply not having an interest in changing anything (Frankfurt 1999c, 104). Therefore, for Frankfurt, this psychic state of satisfaction indicates that there would be no need to continue requiring higher and higher orders of desires in order to determine if something is what you really want, thereby avoiding the infinite regress problem (Frankfurt 1999c, 104).

However, Frankfurt makes an important caveat. Although the state of satisfaction does not require that an agent deliberately act or refrain from acting, there is still an element of reflection that is required for satisfaction. As Frankfurt states,

> It is a matter of simply *having no interest* in making changes. What it requires is that psychic elements of certain kinds *do not occur*. But while the absence of such elements does not require either deliberate action or deliberate restraint, their absence must nonetheless be reflective. In other words, the fact that the person is not moved to change things must derive from his understanding and evaluation of how things are with him. Thus, the essential non-occurrence is neither deliberately contrived nor wantonly unselfconscious. It develops and prevails as an unmanaged consequence of the person's appreciation of his psychic condition. (Frankfurt 1999c, 105)

In a footnote on the same page, Frankfurt compares this state of satisfaction to the state of being relaxed (Frankfurt 1999c). He has a very particular idea of what it means to be relaxed. Although there can be feelings, beliefs, attitudes, or actions that lead to states of relaxation, they do not constitute the state of being relaxed nor are they necessary for relaxation (Frankfurt 1999c, 105). Instead, he defines relaxation as a state of abstaining from worry. Presumably then, Frankfurt is trying to show that in the same way we must stop worrying to be relaxed, we must also abstain from wanting any changes in order to be satisfied.

How does his new conception of satisfaction play out in a typical Frankfurtian hierarchical approach? He uses an example of a person with an addiction again, this time a cigarette smoker instead of a person with a drug addiction (Frankfurt 1999c, 105). Initially, when using the unwilling drug addict example, this problem was presented as the unwilling addict needing to identify with one of their conflicting desires, thereby creating a second-order desire. However, the problem Frankfurt has been contending with is what gives the unwilling addict's desire to not do the drug any authority? How is a higher-order desire supposed to constitute what the unwilling addict really wants? As previously discussed with the infinite regress challenge, we can keep going up in level, requiring a third-order desire, and so on. Here, Frankfurt says, is where satisfaction comes in to solve this problem. As Frankfurt argues in the "The Faintest Passion" (1999), the endorsed higher-order desire to not smoke must be a desire with which the agent is satisfied, and this is what is meant by identifying with the first-order desire. This would mean that the smoker, being satisfied with their desire to not smoke, would not want to change anything and is no longer concerned with any of the conflicting desire options (Frankfurt 1999c, 105). As Frankfurt further explains,

> Hierarchical accounts of the identity of the self do not presume, however, that a person's identification with some desire consists simply in the fact that he has a higher-order desire by which the first desire is endorsed. The endorsing higher-order desire must be, in addition, a desire with which the person is satisfied. And since (as I tried to explain earlier) satisfaction with one psychic element does not require satisfaction with any other, being satisfied with a certain desire does not entail an endless proliferation of higher orders and desires. Identification is constituted neatly by an endorsing higher-order desire with which the person is satisfied. It is possible, of course, for someone to be satisfied with his first-order desires without in any way considering whether to endorse them. In that case, he is identified with those first-order desires. But insofar as his desires are utterly unreflective, he is to that extent not genuinely a person at all. He is merely a wanton. (Frankfurt 1999c, 105)

For Frankfurt, the idea of wholeheartedly deciding on one of your conflicting first-order desires is a stronger position, as well as a more developed explanation than that of his earlier ideas of decisively identifying with a lower-order desire. The choice of lower-order desires, as indicated by one's satisfaction, is now wholeheartedly constitutive of oneself. This intermediate shift for Frankfurt's theory of autonomy, brings him closer to his third and last approach, where one's volitions must come from an essential part of one's character, established by what one cares about.

It has been argued (Velleman 2006; Arpaly 2003) that there is a significant shift from Frankfurt's early work to his later discussions of autonomy (e.g., from his early work "Freedom of the Will and the Concept of a Person" (Frankfurt 1971) to his later work like "Autonomy, Necessity, and Love[4]" (Frankfurt 1999a)). Frankfurt's early work, as previously discussed, endorsed a conception of autonomy in the sense of self-control over one's desires (also termed "agent-autonomy" by Velleman 2006). For Frankfurt, this conception of autonomy required an agent to be self-governed by motives that were internal to the self. Examples used by Frankfurt

[4] "Autonomy, Necessity, and Love" originally published in 1994, reprinted in Frankfurt's collection of essays: Frankfurt, H. G. 1999. *Necessity, Volition, and Love*. Cambridge, UK: Cambridge University Press.

during this time mostly consisted of instances of weakness of will, or *akratic* agents struggling with desire conflicts, where one desire was extremely tempting or irresistible in some sense (e.g., the unwilling drug addict example). In other words, the agents in Frankfurt's examples are ultimately struggling with problems of self-control (e.g., see Arpaly 2003 for detailed discussion of this point).

However, in Frankfurt's more recent work (e.g., "Autonomy, Necessity, and Love" (Frankfurt 1999a) and "Rationality and the Unthinkable" (Frankfurt 1998b)), we can see a shift toward a new conception of autonomy where an autonomous agent's volitions are derived from the essential character of their will. Frankfurt begins to focus more on an agent's character and the volitional unity of the will. This new approach still generally involves self-governance by internal motives, but with a new criterion for internal motives (Velleman 2006). The new criterion is that motives are considered internal to the self when they are "essential to the subject's volitional nature" (Velleman 2006, 332). For Frankfurt, this means that it is these essential motives that must be the source of one's behaviour to be considered autonomous.

We can most clearly see this switch in Frankfurt's approach by looking at the examples he uses (Arpaly 2003, 121). Rather than discussing the unwilling drug addict with problems of self-control, Frankfurt's newer examples examine agents constrained by their deepest values, or their "volitional essences" (Arpaly 2003, 121). For example, Frankfurt discusses a hypothetical military officer, ordered to set off nuclear missiles. The officer is presumed to have voluntarily signed up for the mission, as well as understood the consequences of setting off missiles and the deaths that would follow. However, when it comes time to actually set off the missiles, the officer finds himself unable to follow through with his orders (Frankfurt 1998b, 182). He finds himself unable to go against his deepest values.

Similarly, Frankfurt also gives an example of a man who decides he has reasons to spy on his bride-to-be in order to assess her trustworthiness. However, in the end he also does not go through with the previously decided upon actions because he finds the idea of spying on his future bride revolting (Frankfurt 1998b). Again, this man finds himself unable to go against his deepest values. Frankfurt claims that these examples show that the agents are prevented from acting because they are motivationally constrained by what they care about. Their deepest values do not allow them to act in certain ways. For Frankfurt, this means that these desires are "truly their own" and makes them who they are as a person (Arpaly 2003, 121).

Frankfurt also fleshes out these ideas by making a case for autonomy having its source in "love," as discussed in "Autonomy, Necessity, and Love" (Frankfurt 1999a). For Frankfurt, love is a way to care about things (Frankfurt 1999a, 129), rather than having feelings about something or finding it appealing. As he further explains,

> I shall not attempt to provide a comprehensive analysis of the concept of love. As I shall deploy the concept, it has a very wide scope: love is a species of caring about things, and its possible objects include whatever we may care about in certain ways. Thus loving something—at least, as I propose to construe the matter—is not merely a matter of liking it a great deal or of finding it deeply satisfying, as in "loving" chocolate ice cream or the piano music of Chopin. To love differs from having feelings of a certain type, such as those of powerful attraction or of intense desire or of compelling delight. It is also not equivalent to or entailed by any judgment or appreciation of the inherent value of its object. To love something is quite different from considering it to be especially appealing or precious. The fact that a person recognizes that an object is valuable or that it is good does not imply that he cares about it or, indeed, that he has any particular interest in it at all. (Frankfurt 1999a, 129)

Therefore, love, as construed by Frankfurt, is a way of caring about something that signifies it as an essential part of one's will and thereby internal to oneself. However, Frankfurt also makes an important distinction between active and passive love, with only the former being

central to autonomy. The way he explains the concept of active love contrasts with the Kantian idea of an impersonal, rational autonomy. To briefly summarize the Kantian perspective, Kant famously states that "autonomy of the will is the property of the will by which it is a law to itself (independently of any property of the objects of volition)" (Kant 1998, 47). That is, the will is autonomous when it is acting in accordance with the dictates of rational moral law, rational principles that are independent of the will itself (Kant 1998, 47).

Importantly for Kant, the rational principles of moral law must be universal, and therefore, must also be impersonal. An autonomous agent cannot be motivated by one's own interests, but rather, must act on rational principles that could be universalized. For Kant, autonomy cannot be grounded on personal interests because they are at least partially external to one's will (Frankfurt 1999a, 132). As Frankfurt further explains Kant's position, "to the extent that a person's conduct is guided by his interests, he is being governed by circumstances that do not essentially belong to himself. Therefore, he is necessarily heteronomous" (Frankfurt 1999a, 132). For Kant then, a person must be entirely indifferent to one's personal interests in order to be autonomous, and as Frankfurt states, the commands that the will issues are "issued by no one in particular" (Frankfurt 1999a, 132).

However, Frankfurt strongly disagrees with Kant's position, arguing that it seems reasonable to presume that an autonomous agent, in governing their own actions, would act on things they consider to be of most importance. He finds it implausible that an agent would neglect the things they find important, and further that what they find most important would be following moral law (Frankfurt 1999a, 132). Frankfurt argues that when an agent acts autonomously, their volitions are derived from the essential character of their will. However, unlike Kant, this is not the case when an agent is following impersonal rational moral law, but

rather when a person is acting out of love (Frankfurt 1999a, 132). Love, for Frankfurt is inherently personal. As Frankfurt states,

> Love is, of course, paradigmatically personal. Nevertheless, the unconditional commands of love are not, as Kant suggests, adventitious elements of a person's will. They are essentially integral to it, for what a person loves is a defining element of his volitional nature. When he acts out of love, accordingly, his volitions do derive from the essential character of his will. Thus, the personal grip of love satisfies the conditions for autonomy that Kant believes can be satisfied only by the impersonal constraints of the moral law. (Frankfurt 1999a, 132)

According to this line of argument, for Frankfurt, the distinction between active and passive love has to do with whether a person's interests are conceptually essential or inessential to one's will. If conceptually inessential, then they are considered separable and therefore external to one's will (Frankfurt 1999a, 133). For Frankfurt, if a person's actions are being affected by considerations that are external to their will, then this person is being "acted upon" (i.e., a passive form of love). On the other hand, if one's considerations are essential to one's will, then their will has a hand in actively determining itself (i.e., an active form of love) (Frankfurt 1999a, 133). As he explains himself, "the distinction between heteronomy and autonomy coincides, then, with the distinction between being passive and being active" (Frankfurt 1999a, 133).

As we can see, the requirements for determining if one's desires are internal or external have significantly changed from Frankfurt's earlier work to his later caring approach to autonomy. Instead of relying on identification with one's desires in order to internalize and make them one's own, we have Frankfurt focusing on which considerations are conceptually essential to the volitional character of one's will. Although a motivational source in love conceptually

seems emotional, Frankfurt clarifies that love as the source of autonomy is not actually affective or cognitive. Rather, for Frankfurt, love is strictly "volitional." As he further elaborates,

> Of course, love does ordinarily involve various strong feelings and beliefs that express, reveal, and support it. The heart of love, however, is neither affective nor cognitive. It is volitional. That a person cares about or that he loves something has less to do with how things make him feel, or with his opinions about them, than with the more or less stable motivational structures that shape his preferences and that guide and limit his conduct. What a person loves helps to determine the choices that he makes and the actions that he is eager or unwilling to perform. (Frankfurt 1999a, 129)

Normally, when we think of love, we would assume that this is a paradigmatically affective concept. As Frankfurt mentions in the quote above, love does usually involve strong feelings. However, love, as Frankfurt means it to be used here, is a type of motivational structure. Love, or the things you care about, act as a constraint on your choices that helps keep your choices in line with your will. What a person loves is considered a part of one's essential volitional character. Therefore, love can provide the required agential authority to our actions. This is why Frankfurt claims that love is neither cognitive nor affective, but rather volitional. The feelings and beliefs that accompany feelings of love are not central to autonomy, but rather the motivational constraint that comes from love is what is fundamental to acting autonomously. Furthermore, the fact that the motivational constraint comes from the essential parts of one's will, maintains the necessary authority over our actions.

Frankfurt's care approach to autonomy is rooted in the Humean idea of the importance of our essential character for morality. For Hume, we cannot hold people morally responsible unless we can infer their character from their "internal, and consistent, motives" (Russell 2007). Importantly for Hume, these consistent character traits give rise to our moral sentiments (Russell 2007). Hume argued that intentional actions are the product of the passions (the direct passions

in particular, such as desire, aversion, hope, fear, joy, grief, as well as our instincts, such as hunger, lust, etc.), and that these motivating passions are caused by ideas of pleasure and pain (Cohon 2004).

Although Hume holds that the passions are the only motivating source of intentional action, he does not claim that reason has no role at all. He argues that reason can provide information and helps us learn from previous actions. However, reason alone cannot move us to act. As Hume himself famously stated, "reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (Hume 1985, 462).

However, Hume also points out an interesting implication to his argument that only the passions can be a source of action. If this is the case, then there cannot be a battle between a reason-based action and a passion-based action because actions can never have their source in reason. These kinds of battles have been central to debates about the will, and how we define self-control. As Hume states,

> Since reason alone can never produce any action, or give rise to volition, I infer, that the same faculty is as incapable of preventing volition, or of disputing the preference with any passion or emotion…Nothing can oppose or retard the impulse of passion, but a contrary impulse; and if this contrary impulse ever arises from reason, that latter faculty must have an original influence on the will, and must be able to cause, as well as hinder any act of volition. But if reason has no original influence, 'tis impossible it can withstand any principle, which has such an efficacy, or ever keep the mind in suspence a moment. Thus it appears, that the principle, which opposes out passion, cannot be the same with reason, and is only call'd so in an improper sense. (Hume 1985, 462)

From a Humean perspective, then, this would mean that we have been thinking about paradigmatic self-control dilemmas in the wrong way. That is, typical examples of an *akratic* agent, where emotional impulses tempt us away from our reasoned judgments, are not a battle

between reason-based judgments and emotion-based impulses. For Hume, the temptation and our willed choice both have their sources in the passions—both choice options are emotional.

Thus, Hume's position is that only the passions can be a source of action, and the motivating passions (including instincts) are caused by ideas of pleasure and pain. Additionally, one's essential, internal, and constant character traits give rise to our moral sentiments, a requirement for moral responsibility. Frankfurt, broadly following in the tradition of Hume, also argues that autonomous action must come from an internal "essential, volitional character," and that desires that stem from our essential character are granted full agential authority. In fact, Frankfurt is often understood as the more affective approach to autonomy because of his focus on desires and the motivational structure of the will, in contrast to the more rational, Kantian approaches (Tappolet 2014, 165).

However, for Frankfurt the motivational source of autonomous action is not the passions generally, as Hume would argue. Frankfurt takes a much narrower approach to the motivating source of autonomous action than Hume's position. Frankfurt specifies that it is only active love, not any other emotions, that can be an authoritative source of action. As Frankfurt further explains,

> However imposing or intense the motivational *power* that the passions mobilize may be, the passions have no inherent motivational *authority*. In fact, the passions do not really make any *claims* upon us at all. Considered strictly in themselves, apart from whatever additional impetus or facilitation we ourselves may provide by acceding to them, their effectiveness in moving us is entirely a matter of *sheer brute force*. There is nothing in them other than the magnitude of this force that requires us, or that even encourages us, to act as they command. (Frankfurt 1999a, 137)

For Frankfurt, there is a clear distinction between being moved by our emotions by "sheer brute force" and being constrained by what we care about—or as he says, the fundamental

difference between "being 'captivated' by love and being 'enslaved' by passion" (Frankfurt 1999a, 137). Frankfurt acknowledges that other passions and impulses can forcefully move us to action, but he treats these similarly to how he treats compulsions. That is, the passions, like compulsions, can be too powerful to resist and force us to act, but they are independent of our own desires and intentions (Frankfurt 1999a, 136). In this sense, Frankfurt claims the passions are alien to us and therefore do not have motivational authority (Frankfurt 1999a, 136). Active love, on the other hand, is what constitutes our essential nature as a person, and thereby helps define our will itself by constraining our actions according to what we care about. Frankfurt further explains this idea:

> These constraints cannot be determined by conceptual or logical analysis. They are substantive rather than merely formal. They pertain to the purposes, the preferences, and the other personal characteristics that the individual cannot help having and that effectively determine the activities of his will. In other words, they are specified for any given person by what he loves. Our essential natures as individuals are constituted, accordingly, by what we cannot help caring about. The necessities of love, and their relative order or intensity, define our volitional boundaries. They mark our volitional limits, and thus they delineate our shapes as persons. (Frankfurt 1999a, 138)

Therefore, for Frankfurt, the authority that love has over our actions can be derived from one's own essential nature as a person (Frankfurt 1999a, 138). This provides Frankfurt with a way to directly tie together our essential volitional characteristics with our actions and finally provide an answer to how we know our desires are truly what we really want.

Despite the influence of Hume on Frankfurt's approach, there is a fundamental difference between the two approaches in the way they view emotions being a motivational source of autonomous action. For Hume, emotions are the only motivational source of autonomous actions. For Frankfurt, emotions can powerfully motivate actions by sheer brute force but cannot

motivate autonomous actions because they cannot provide the required agential authority. Only active love can motivate autonomous actions.

This key difference between Hume and Frankfurt accounts for the differences in how they view the conflict between reason and emotion. For Hume, there can be no competition between a reason-based source of action versus an emotion-based source of action, the kind of self-control battle that plays a large role in Frankfurt's early and intermediate work. Hume argues that reason cannot be a source of autonomous action, only the passions can motivate our actions. And while Frankfurt's theory of autonomy is largely known as an affective, desire-based account, he dismisses emotions as an authoritative motivational source. Instead, he only focuses on active love as a motivational constraint—even dismissing the subjective feelings and beliefs that come along with love. By Frankfurt dismissing emotions as a source of autonomous action and reiterating the idea that emotions can only move us by "brute force," he maintains the historical notion that emotions act as distractions to our actual willed desires. This perspective on emotions certainly pushes Frankfurt further away from the Humean tradition of prioritizing emotions as the source of our autonomous actions. However, within the autonomy literature, Frankfurt remains the Humean example of a theory of autonomy.

1.2 Watson's Structural, Reasons-Responsive Theory of Autonomy: A Platonic Approach

In "Free Agency" (Watson 1975), Gary Watson provides another version of a structural theory of autonomy. Like Frankfurt, Watson's theory focuses on the structure of an agent's will, but he approaches it from a Platonic, rather than a Humean perspective. Watson's approach to autonomy is based on the Platonic idea of the division of the soul into the rational, spirited, and appetitive parts, each providing independent sources of motivation for an agent. The ideal, well-

ordered soul for Plato requires that the rational part of the soul governs the spirited and appetitive parts. Similarly, for Watson, an agent's actions are free when the rational part of the soul is the source of motivation for one's actions.

Following Plato, Watson argues that we can distinguish rational sources of motivation from appetitive and spirited sources. This forms the foundation for his theory in which he distinguishes between "desiring" and "valuing," as irrational and rational sources of motivation, respectively (Watson 2004, 25)[5]. As previously discussed, in response to Frankfurt's early work on the motivational structure of an agent's desires (Frankfurt 1971), Watson questions why we should privilege higher-order desires, and whether higher-order desires can be an effective measure in determining what an agent really wants. He argues that the idea of higher-order desires cannot do the work that Frankfurt wants it to do—that is, it cannot help us differentiate between desires that are more or less truly our own (Watson 2004, 28). As Watson sees it, agents "do not (or need not usually) ask themselves which of their desires they want to be effective in action; they ask themselves which course of action is most worth pursuing" (Watson 2004, 30). Therefore, for Watson, the crucially important step in determining autonomy is to differentiate the motivational sources of our desires—do they stem from the rational or irrational parts of ourselves? When the source of the desire is rational then the course of action is worth pursuing and considered of value.

For Watson, values are essentially related to judging something good and provides reasons for actions that arise from the "rational" self. This approach makes reason the original source of autonomous action, therefore providing authority over one's choice. In contrast, we

[5] Gary Watson's "Free Agency" (1975), reprinted in Watson, G. 2004. *Agency and Answerability: Selected Essays*. New York: Oxford University Press.

merely desire objects or states of affairs that have an irrational source of motivation (Watson 2004, 25). Following this distinction between values and desires, Watson distinguishes between the valuational system and motivational system. When there is harmony between these two systems, Watson considers the agent's actions free (Watson 2004, 25).

For Watson, what makes an individual a free agent is their capacity for critical evaluation, a capacity he calls "normative competence" (Watson 2004, 2). An agent's normative competence allows them to identify the sources of their motivational desires. Consider an example Watson gives in "Free Agency" (Watson 2004). Here Watson discusses in detail the example of the desire to eat and the ways in which we can determine the source of the agent's wants and desires. As Watson states,

> Obviously, to identify a desire or want simply in terms of its content is not to identify its source(s). It does not follow from my wanting to eat that I am hungry. I may want to eat because I want to be well-nourished; or because I am hungry; or because eating is a pleasant activity. This single desire may have three independent sources. (These sources may not be altogether independent. It may be that eating is pleasurable only because I have appetites for food). (Watson 2004, 21)

For Watson, this means we can have wants and desires that stem from independent sources of motivation. Wanting food can have its source in the rational, valuational part by wanting to be well-nourished. On the other hand, desiring food can have its source in the irrational appetitive part, if one's motivation is purely pleasure (Watson 2004, 21). However, at times this can be complicated by overlap or conflicts between motivational sources. In fact, Watson argues that these sorts of conflicts exist *because* of these independent sources of motivation. Since desires can arise that are independent of your valuational system, or personal judgments of the "good," we can end up with conflicts between what is valued and what is merely desired. Watson claims that it is an essential feature of the appetites and passions that

they consist of desires that are independent of a person's judgment of the good (Watson 2004, 21).

Although Watson generally considers the passions as falling under the irrational category, this does not apply to all emotions. Certain emotions, such as lust, may have an appetitive source. However, Watson points out that many other emotions, such as resentment for example, involve beliefs and some sort of valuation (Watson 2004, 21n6). In these cases, he suggests the source of motivation would be from the *Thumos*, or spirited part—a part of the soul that is, as he says, "partly rational" (Watson 2004, 21n6).

For Watson, it is important that the valuational system must have "a considerable grip" on our motivational system (Watson 2004, 26), and this is the case even when there are motivational factors at play other than the valuation motivations. Watson quotes a key section in Plato's *Phaedrus*, which highlights the issues that arise in these types of conflicts related to the battle between rational and irrational sources of motivation:

> To quote just one of many suggestive passages: "We must . . . observe that within each one of us there are two sorts of ruling or guiding principle that we follow. One is an innate desire for pleasure, the other an acquired judgment that aims at what is best. Sometimes these internal guides are in accord, sometimes at variance; now one gains the mastery, now the other. And when judgment guides us rationally toward what is best, and has the mastery, that mastery is called temperance, but when desire drags us irrationally toward pleasure, and has come to rule within us, the name given to that rule is wantonness" (*Phaedrus*, 237e-238e). (Watson 2004, 17n3)

Similar to Plato's position in *Phaedrus*, Watson argues that a free agent is one that acts on their valuational motives, rather than their irrational desires. However, Watson does address this apparent dualism between the rational and irrational parts, arguing that we should not necessarily consider it two antagonistic halves of our self (Watson 2004, 22). To this point, he says that the relationship between evaluation and motivation is intricate, and there may be many

activities that we deem worthwhile, despite their motivational source being appetitive (Watson 2004, 22). However, the very fact that we can have motivations that are independent of our values does mean that there is the possibility that an agent could be motivated to do things that they do not consider worthwhile. This possibility leads to the potential problem of an agent being obstructed by their own will, that as Watson says, is the basis for the fundamental problem of free action (Watson 2004, 22).

In "Free Agency" (Watson 1975), Watson highlights the essential differences between a Humean approach, like Frankfurt's theory of autonomy, and his own Platonic approach to autonomy. The fundamental differences underlying Frankfurt and Watson's strucutural approaches to autonomy relate back to the key differences between Platonic and Humean approaches to understanding reason and emotion. For instance, as discussed by Watson, Plato distinguished between rational and irrational parts of the soul—between reason and appetite (Watson 2004, 15). Hume had a similar distinction between reason and the passions. However, as Watson summarizes the Humean viewpoint,

> Reason is not a source of motivation, but a faculty of determining what is true and what is false, a faculty concerned solely with 'matters of fact' and 'relations among ideas'…A Humean might allow Reason a crucial role in deliberation. But it's essential role would not be to supply motivation—Reason is not that kind of thing—but rather to calculate, with a context of desires and ends, how to fulfill those desires and serve those ends. (Watson 2004, 15)

However, for Plato, reason is not a calculating mechanism, but rather, it is itself the source of motivation, where "the desires of Reason are desires for 'the Good'" (Watson 2004, 16). That is, the rational part of the soul is what helps us determine which actions have value and how much value, therefore, according to Watson, "values provide reasons for action" (Watson 2004, 17). This is not only fundamentally different than the Humean understanding of reason,

but also the Humean understanding of emotion. From a Platonic approach, reason must be in control of our emotions; yet, from the Humean approach, reason ought to be "slave to the passions," as discussed in the previous section.

Hume's approach to reason and emotion as motivational sources of action is less combative than Plato's—if reason cannot be a source of motivation, than it simply does not compete with passion-based sources of motivation. However, Hume does still distinctly separate the two concepts, and positions reason in a role to be governed by the passions. On the other hand, from a Platonic perspective, both reason and emotion can be distinct sources of motivation that often compete, but only reason can be a source of autonomous action. It is from these early ideas that we get the notion that there can be rational and irrational (emotional) sources of motivation that compete for dominance, and only the rational source can pursue the good. This combative perspective of reason versus emotion still resonates within self-control and autonomy literature today, as will be discussed in later chapters.

1.3 Korsgaard's Reasons-Responsive Theory of Autonomy: A Kantian Approach

The preceding sections have focused on two different versions of structural theories of autonomy: Frankfurt's Humean approach, and Watson's Platonic approach. Although Watson's Platonic approach is also considered a reasons-responsive account, the focus of this section are reasons-responsive approaches to autonomy that are not structural. That is, I will consider reasons-responsive accounts that do not focus on an agent's internal motivational structure when determining a capacity for autonomy. Instead, the non-structural reasons-responsive accounts insist on the external condition of the agent being responsive to reasons for their actions. From

this approach, an agent governs herself if her motives, or the mental processes that produce those motives, are responsive to a wide range of reasons for acting (Buss and Westlund 2018).

A prominent method for how this type of responsiveness to reason can be understood is through a Kantian framework. For Kant, our capacity for autonomy is based on our ability to impose universal moral law upon ourselves (Christman 2003). According to Kant, we can use reason to identify actions that weigh upon us as categorical imperatives, ensuring that our actions conform to these maxims. This ability to use reason and self-legislate is the foundation of human autonomy for Kant.

An influential Kantian reasons-responsive account is Christine Korsgaard's theory of autonomy (Korsgaard 1996; 2009). Following Kant, Korsgaard argues that our capacity for reflection is central to being able to use universal law to govern our choices and actions. She claims that the "reflective structure of human consciousness requires that you identify yourself with some law that will govern your actions and it also requires you to be a law to yourself" (Korsgaard 1996, 104). For Korsgaard, autonomy is the source of obligation (Korsgaard 1996, 104). However, unlike Kant, Korsgaard claims that we have many different "practical identities," which she states can be best understood as "a description under which you value yourself and find life worth living and your actions worth undertaking" (Korsgaard 1996, 101). For Korsgaard, our practical identities are supposed to capture the "you" that is making the decisions. As she further explains,

> When you deliberate, it is as if there were something over and above all of your desires, something which is you, and which chooses which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as being expressive of yourself. (Korsgaard 1996, 100)

For Korsgaard, we all have many different self-concepts (e.g., thinking of ourselves as a mother, friend, teacher, a citizen of a particular nation, etc.), which gives rise to many different reasons for acting and obligations. The conceptions that are most important to us are the ones that give rise to unconditional obligations. According to Korsgaard, to violate these unconditional obligations is to lose your integrity and no longer be who you are (Korsgaard 1996, 102).

Central to her theory of autonomy are two main claims. First, following the Kantian tradition, Korsgaard claims that the reflective structure of human consciousness gives us authority over ourselves. Reflection creates the distance from our desires and impulses, which is required in order for us to be able to make laws for ourselves. According to Korsgaard, to make a law for yourself is at the same time to give expression to a practical conception of your identity and these practical identities determine what will count as reasons for action (Korsgaard 1996, 129). Second, reflection helps us recognize our identity as humans, which Korsgaard defines as "reflective animals who need to have practical conceptions of our identity in order to act and live" (Korsgaard 1996, 129). She argues that treating your human identity as a source of reasons and obligations is to have what she calls a "moral identity."

According to Korsgaard's view, what makes human action special is our capacity for reason. Reason, for Korsgaard, lies in the particular way that we are self-conscious as humans. We are aware of the fact that we have certain desires or emotions, as well as being aware of our inclinations to act based on those desires or emotions. That is, we are aware of our potential grounds for action (Korsgaard 2009, 115). She claims that this puts us in a unique position, "once we are aware that we are inclined to act in a certain way on the ground of a certain incentive, we are faced with a decision—should we do it?" (Korsgaard 2009, 115). The fact that

we can ask ourselves this question makes it possible for us to be rational, autonomous agents. This specific type of self-consciousness opens up what she calls a "reflective distance" between our motivations and our response.

In order to support her theory of autonomy, Korsgaard differentiates between two types of autonomous action. The first considers how a non-human animal might govern their own actions. In this case, she claims that the movements of non-human animals are governed by a type of principle—that is, their actions are governed by the animal's instincts. Korsgaard argues that their instincts, in that they spring from the animal's own nature, can be considered to be "the animal's will, the laws of their own causality" (Korsgaard 2009, 104). The animal's instincts determine what the animal does, and by the animal acting according to their own laws of causality, they are therefore autonomous in a way.

However, this is not the same type of autonomy that a human is capable of achieving. This, Korsgaard claims, is a deeper sense of autonomy where not only are we guided by the laws of our causality, but we also go one step further and choose *our own* laws or principles from which our actions are to be guided. A non-human animal does not choose the content of their instincts, like humans are able to choose the content of our action-guiding principles (Korsgaard 2009, 108).

Korsgaard explains how it is that humans go about choosing their principles. Following Kant, Korsgaard claims that a human's starting point for action comes from what Kant called our "incentives" (Korsgaard 2009, 105). Korsgaard defines incentives as "a motivationally loaded representation of an object" (Korsgaard 2009, 105). By object, she means any substances, states of affairs, or activities. When we are aware of a feature of an object being either appealing or aversive, this is us being subject to an incentive (Korsgaard 2009, 105).

Importantly for Korsgaard's argument, an incentive does not directly cause our movements. It must work in conjunction with our principles (Korsgaard 2009, 120). It is our principles that make us able to determine if we should act in response to the incentive. She argues that it is our self-consciousness that transforms incentives to inclinations, inclinations being our reaction to the incentive (Korsgaard 2009, 122). However, the reasons to act do not come from the desired object, rather they come from our own response to that object. It is our choices that confer the value on the objects, even when these choices are responding to features of the objects (Korsgaard 2009, 123).

And where do these incentives first come from? Korsgaard claims that many of our incentives come from our own various practical identities that we have based on the many roles and relationships that we have and value in our lives (Korsgaard 2009, 126). However, for our actions to be autonomous, these various practical identities must be unified as a whole.

How this unification is accomplished is a part of Korsgaard's constitutional model of deliberative action. Her main claim is that your identity is constituted by your choices or actions (Korsgaard 2009, 19). This identity is necessary in order for an agent's actions to be attributed to themselves. Korsgaard takes us through arguments for how a person can and must unify themselves in order to act—a process she calls deliberation—from both a Kantian and Platonic perspective. She presents two connected arguments to establish her position.

First, Korsgaard argues that to be an agent capable of action, one must be the autonomous and efficacious cause of her own actions. Being efficacious would necessitate that an agent is successful in bringing about whatever state of affairs she intended through her action (Korsgaard 2009, 82). To be autonomous, the agent's movements must be self-determined (Korsgaard 2009, 83). She argues that these two essential properties of an agent correspond to Kant's two

imperatives: the hypothetical imperative and categorical imperative (Korsgaard 2009, 83). For Korsgaard, the hypothetical and categorical imperatives are constitutive principles of action that she claims picks out two crucially important features of action (Korsgaard 2009, 72). First, "if you will an end, you must will the means to that end" (Korsgaard 2009, 68), and second, a "person who wills an end, constitutes himself as the cause of that end" (Korsgaard 2009, 68).

Therefore, Korsgaard claims that the hypothetical imperative picks out the *cause* aspect of action—that "to act, is to constitute yourself as the cause of an end" (Korsgaard 2009, 72). The categorical imperative, on the other hand, highlights the other feature of action—that you make *yourself* the cause of the action (Korsgaard 2009, 72). It is here that Korsgaard is trying to determine the "you" behind your actions, for she claims that when you deliberate "it is as if there is something over and above all your incentives… something which is you, and which chooses which incentive to act on" (Korsgaard 2009, 72).

Korsgaard claims the Kantian position would mean that in order for you to determine your own causality it must come from your unified whole identity, and for this, you must will your maxims as universal laws (Korsgaard 2009, 72). As Korsgaard explains, "this means that when you determine yourself to be the cause of the movements which constitute your action, you must identify yourself with the principle of choice on which you act" (Korsgaard 2009, 75). However, she also clarifies that it is not that you are identifying directly with a rational principle. Rather, it is your rational constitution that you are identifying with and your constitution is the one that decides on the rational law that governs your actions—or that reason should govern generally (Korsgaard 2009, 135).

The Platonic argument she also uses works similarly to Kant's, where we can understand Plato's principles of a just soul as also being constitutive principles of action in that they unify

the person acting, making action possible (Korsgaard 2009, 158). As Korsgaard claims, "for Kant, like Plato, reason must rule for the good of the whole" (Korsgaard 2009, 158). For this, like Watson, Korsgaard goes back to Plato's argument of the tripartite soul. She argues that the three parts of the soul (appetite, spirit, and reason) correspond to the three parts of deliberative action.

For Korsgaard, deliberative action begins with our desires, or appetites. We are conscious of these desires, and they propose to us a possible action to acquire a particular end. However, because we are conscious of these desires and this corresponding proposal, we do not have to automatically act to satisfy these desires. We can decide whether or not we should satisfy the desires (Korsgaard 2009, 141). How does this line up with Plato's tripartite soul? Korsgaard claims that we can think of deliberative action in the following steps: appetite makes a proposal, reason decides whether or not to act on this proposal, and spirit carries out the decision (Korsgaard 2009, 141).

For a well-ordered soul, the appetitive part is controlled by the "desires of reason," which aims for the good of the whole, unified person (Korsgaard 2009, 146). Korsgaard specifies here, a very particular type of control that she has in mind. Ideally it is not reason forcefully repressing or constraining our appetites, but rather all parts of the soul agreeing that reason should rule all the other parts. Korsgaard claims that our desires, or appetites, would not on their own be able to be aimed toward the good. Rather, appetitive desires are aimed at particular objects, whereas reason's desires are aimed at the good of the whole person (Korsgaard 2009, 147). When all of these parts work together harmoniously, an agent is able to act from a unified self which is required for action in itself (Korsgaard 2009, 148). As Korsgaard says, "desires and impulses

may operate within the unjust person, but the unjust person is completely incapable of accomplishing anything because the unjust person cannot act at all" (Korsgaard 2009, 152).

In sum, Korsgaard's constitutional model begins with a rational agent that is aware of their desires and the potential grounds of their actions. This awareness, or self-consciousness, gives humans the reflective distance required in order to be in control of these desires, where we do not need to automatically act on these desires in the way that non-human animals must act on their instincts (Korsgaard 2009, 213). Korsgaard calls the act of deciding whether or not to act on a desire "reason," where you must unify the parts of yourself by making a choice (Korsgaard 2009, 213). Your choice requires a principle, the specified constitutive principles, where your actions have to come from your constitutional rule over yourself (Korsgaard 2009, 213). These constitutive principles come from the categorical and hypothetical imperatives, as well as Plato's principles of a just, well-ordered soul. Deliberation, for Korsgaard, is this act of unifying yourself behind a set of movements in order to count these movements as your own, as movements of a rational agent (Korsgaard 2009, 213). Otherwise, Korsgaard claims, we are "merely a heap of impulses, and not an agent after all" (Korsgaard 2009, 213).

Although Korsgaard uses Plato's tripartite soul to explain how an agent is supposed to harmonize the rational, emotional, and motivational parts of themselves—a Kantian framework is used to determine how one actually makes and carries out autonomous decisions. There are two key points for determining one's autonomy from a Kantian perspective that we can see underlying Korsgaard's approach to autonomy. First, emotions cannot lead us to the moral law. We must use reason to determine our principles, and our actions must come from those principles. We can see this in Korsgaard's model of deliberative action, where our choices

require normative principles, principles that can only be determined by reason. As Korsgaard clarifies,

> …reason is a power we have in virtue of a certain type of self-consciousness—consciousness of the grounds of our own beliefs and actions. This form of self-consciousness gives us a capacity to control and direct our beliefs and actions that the other animals lack, and makes us active in a way that they are not. But it also gives us a problem that the other animals do not face—the problem of deciding what to count as a reason for belief or action. To put the point another way, this form of self-consciousness makes it necessary to take control of our beliefs and actions, but we must then work out how to do that: we must find normative principles, laws, to govern what we believe and do. The distinctive feature of human beings, reason, is therefore the capacity for normative self-government. (Korsgaard 2009, xi)

The second key point from a Kantian perspective is the idea that emotions can interfere with our practical reasoning. To avoid this problem, in addition to her requirement for reason-based normative principles, Korsgaard also uses the Platonic model of the soul to argue that an autonomous person must be democratically ruled by reason, not appetite or spirit. This type of rational self-control for Korsgaard, ensures that an agent is acting on desires aimed at the good of the whole person—thereby avoiding acting on emotional impulses. As the name suggests, reasons-responsive accounts of autonomy prioritize reason as the only place from which our decisions and actions can originate in order to be autonomous, with strong rational influences from both Plato and Kant. Korsgaard follows in this tradition, using elements of both Plato and Kant to support her theory of autonomy.

Conclusion

Classic contemporary theories that have been highly influential within the autonomy literature, such as the theories of Frankfurt, Watson, and Korsgaard, have provided vastly different approaches to the central questions raised in relation to individual autonomy: how do we determine if what we want is what we *really* want, how is this self-governing authority granted, and a question of central importance to this chapter, is our motivational source of action based in reason or emotion?

Frankfurt, Watson, and Korsgaard each position themselves according to the historical tradition they generally follow: Hume, Plato, or Kant. The aim of this chapter was to highlight some of these traditional influences and the way in which they have impacted contemporary approaches to autonomy. These early influences have had a wide impact, affecting how we understand the source of motivation, our understanding of self-control, and how agential authority over our actions ought to be determined. Frankfurt is primarily concerned with the motivational structure of an agent, and whether the agent's desires are constrained by what they care about. Watson, on the other hand, highlights the importance of an agent's motivation being sourced in the "rational." And lastly, Korsgaard's account requires that an agent be in control of their spirited and appetitive parts in order to be properly responsive to rational principles. The rational principles are required for the agent to be able to deliberate over their desires.

For Frankfurt, being the least "rational" in the group, we would have expected to see a stronger role for emotion in his theory of autonomy. However, as I argued above, Frankfurt generally dismisses emotions as being an authoritative source of autonomy. Instead, he focuses on a volitional construct of "active love," which helps constrain our actions with what we care about. His approach is considered an example of an affective account of autonomy in that it is

rooted in the motivational structure of our desires. However, he does clarify that he construes active love as strictly volitional, rather than affective or cognitive.

Watson's approach, while similarly structural, follows in the Platonic tradition of identifying and separating the sources of motivational desires into rational and irrational categories. The rational sources of motivation have the authority to be self-legislative in that they are deemed to be pursuing "the good." For Watson, it is not a matter of levels of desires (lower- and higher-order desires), but rather that fundamentally some desires are rational and are therefore representative of what we value while others are irrational and therefore only representative of our more impulsive desires. Naturally then, to be considered autonomous, Watson claims that we must follow the rational desires that lead to the good.

And lastly, I examined Korsgaard's Kantian/Platonic approach to autonomy. For Korsgaard, a rational agent is at the centre of her constitutional model. To be "in control," the rational agent must be aware of the motivational sources of their desires and be able to deliberate on the possibility of acting on these desires. For Korsgaard, this moment of reflecting on our motivational urges is what she calls "reason," where an agent unifies themselves by making a choice, a choice that requires principles of reason. Within this approach we can see a specific type of self-control that is required for autonomy, where an agent must be able to use reason to control one's emotions and impulses. She claims that we must reflect on our desires and use reasoned principles in order to make our choice. Korsgaard's Kantian approach to autonomy, like Watson's Platonic approach, ultimately prioritizes reason as the necessary requirement for agential authority over our autonomous actions.

Therefore, even in the more contemporary approaches to personal autonomy, we can see the lingering effects of the historical conceptions of emotion and reason competing for control

over our autonomous actions. All three contemporary approaches to autonomy—Frankfurt, Watson, and Korsgaard—are generally dismissive of the input of emotions into autonomous decision-making and their ideas of self-control over our desires. The next two chapters will turn to the neuroscience of both decision-making and self-control to see what role, if any, emotions do play in our decision making and self-control processes. I will later return to Frankfurt and Korsgaard to further discuss whether their approaches are compatible with the contemporary neuroscience research on decision-making, control, and the conscious feelings of agency.

Chapter 2

The Neuroscience of Decision Making: Reward, Motivation, and Value

Introduction

As discussed in Chapter One, an important component of personal autonomy is decision-making. Therefore, it is important to consider the research pertaining to how we make decisions and how our motivational system supports this process. Given that the autonomy literature often raises questions about the impact that reason and emotion have on our decisions, this is an important question to address in this chapter as well. The neuroscientific research on decision-making suggests that affect plays a central role in making decisions, as will be discussed in detail in this chapter.

This chapter reviews the decision-making research and examines how cognitive, emotional, and motivational information is integrated in the decision-making process. I argue that frameworks that pit reason against emotion in decision-making, like dual-process models, are no longer the consensus. Instead, the empirical research has moved toward an understanding of decision-making as an integrated decision-making system, encompassing widely distributed areas of the brain (both cortical and subcortical), one that crucially involves cognition, emotions, and motivational processes like the reward network. This chapter examines the neuroscience underlying decision-making, the role that the motivational/reward system plays in this process and highlight some of the central regions/circuits that are important for computing value.

2.1 Perceptual Decision-Making Versus Value-Based Decision-Making

The decisions we make every day can take many forms, ranging from simple choices like deciding between eating an apple or an orange, to complex, strategically planned life decisions, such as choosing a future career. Given the complexity of the topic, researchers have focused on different aspects of the decision-making process, resulting in a vast array of research on how we make decisions. This section will review some of the prominent lines of research, highlighting the progress thus far in understanding the neural correlates of decision-making. I examine two central components that influence our decisions: the perceptual salience of stimuli and the subjective value of the stimuli (Lou et al. 2015). I consider each of these dimensions of decision-making, first looking at perceptual decision-making, then moving on to value-based decision-making.

2.1.1 Perceptual Decision-Making

Toward the end of the 20th century, huge strides were made in the understanding of the sensory and motor systems in the mammalian brain (Glimcher 2003). The sensory processes were identified by key experiments (Kuffler 1953; Mountcastle 1957; Hubel and Wiesel 1959), which found that individual cortical neurons could be viewed as highly-specific sensory receptors, organized in groups of neurons with related sensory properties (Glimcher 2003, 138). In addition to the work being done on mammalian sensory processes, major advancements were being made on movement control in mammals. A particular focus of the experiments was a type of eye movement, referred to as saccades. Saccades are rapid shifts in the line of sight, and researchers focused on understanding how these types of eye movements were produced (Glimcher 2003, 138).

Research by Wurtz and Goldberg (Wurtz and Goldberg 1972) and Bruce and Goldberg (Bruce and Goldberg 1985), among others, showed that neurons in areas like the cortical frontal eye fields and the mesencephalic superior colliculus mapped representations for saccadic eye movements (Glimcher 2003, 138). With these cortical and subcortical mappings encoding saccadic eye movements, it became possible for researchers to use the saccadic eye movement system as the primary model for understanding the connections between specific sensory receptors and specific motor neurons (Glimcher 2003, 138–39).

This link is crucial for a full understanding of perceptual decision-making, defined as the "process of choosing an appropriate motor action based on perceived sensory information" (Shushruth et al. 2018, 6350). Consequently, a few different research groups (Newsome et al. 1989; Glimcher and Sparks 1992; Hanes et al. 1995) began to study perceptual judgments through the kinds of decision-making processes where a sensory stimulus triggers one of two possible saccades, or rapid eye movements (Glimcher 2003, 139).

Newsome et al. (Newsome et al. 1989) began experimenting with the visual-motion sensitive neurons of the primate middle temporal area (area MT), looking into whether the activity of neurons in area MT was reflected in the perceptual experience of determining the direction (leftward or rightward saccades) of visual motion (Glimcher 2003, 139). The working hypothesis of Newsome et al. (Newsome et al. 1989) was that the activity of MT neurons could predict the eye movements when presented with the motion stimulus, which their experiments were able to confirm (Glimcher 2003, 141). Their results suggested that the sensory neurons in area MT could be considered to be a set of receptors that were each capable of identifying a particular stimulus (Glimcher 2003, 141). With the sensory neurons acting as receptors, it

suggested the possibility that the outputs of these neurons could then be passed to a simple circuit, leading to the activation of either leftward or rightward saccades (Glimcher 2003, 141).

In order to test their hypothesis that a simple circuit might exist, Shadlen and Newsome (Shadlen and Newsome 1996) created a model that could account for the integrative elements needed between the sensory and motor elements of this simple circuit. Such an integrative element would describe the process for transforming sensory signals in area MT into inputs for the superior colliculus and the frontal eye fields (the areas where the saccade was triggered) (Glimcher 2003, 141).

Shadlen and Newsome's work (Shadlen and Newsome 1996; 2001) focused on one particular area—the neurons in the lateral intraparietal area (area LIP)—as a possible site for this integrative element because it was already known to have connections to the frontal eye fields and visual areas like area MT (Glimcher 2003, 144). The results of their experiments in area LIP confirmed their hypothetical model, showing that neurons in area LIP likely acted as an integrator of sensory information, biasing sensorimotor decisions (Glimcher 2003, 144–45).

At the same time, Schall and colleagues (Hanes et al. 1995; Hanes and Schall 1996; Thompson et al. 1996; 1997) were working on the next step, trying to figure out how sensory or integrative signals actually trigger a saccadic eye movement (Glimcher 2003, 145). Their experiments focused on decisions that were to be made as quickly as possible, therefore using the timing of the actual eye movement in order to identify the point in time where the neural circuitry in the eye fields was committed to making the particular eye movement (Glimcher 2003, 146). Their studies suggested that,

> [A]ctivity in the frontal eye field, perhaps shaped by activity in areas like LIP, might be driven toward a discrete biophysical threshold while the decision process is underway. Crossing that biophysical threshold, which might shift the frontal eye field neurons themselves (or the closely related neurons of the

> superior colliculus) into a higher firing rate (or burst) mode, might then serve as the physical trigger that executes the selected saccade. (Glimcher 2003, 147)

These results began to come together, providing a picture of what sensory-driven decision-making might look like. According to Glimcher's 2003 review, a general outline of the process of simple sensorimotor decisions involve neural receptors in the visual cortices, which in turn passes these sensory signals on to the neurons in area LIP of the parietal cortex, which then combines or integrates the information (Glimcher 2003). This signal then gets passed on to the frontal eye fields. The movement occurs when, at this point, neuron activity in the frontal eye fields crosses a biophysical threshold (either in the cortex or colliculus) (Glimcher 2003, 148). In sum, we can think of perceptual decisions as the "accumulation of evidence over time, for and against possible alternatives, until a threshold level is reached and a response is executed" (Gold and Shadlen 2007).

Interestingly, what also came out of these experiments was the observation that this simple sensorimotor circuit may not just be processing sensorimotor decisions, but may also be participating in nonsensory decision-making (Glimcher 2003, 153). In order to further test this possibility, Platt and Glimcher (Platt and Glimcher 1999) began experiments in area LIP in order to determine if changes in prior probabilities and in the value of particular saccade choices influenced the neurons in area LIP (Glimcher 2003, 153). What they found was that neurons in area LIP were "encoding something more like the value of each movement, relative to the available alternatives, than the absolute value of each movement" (Glimcher 2003, 153). This would suggest that sensory signals are analyzed and then combined with nonsensory signals that encode value information, such as the prior probability that each movement may be reinforced as well as the value of the outcomes of each movement (Glimcher 2003, 157). This activity appears to be represented in the "posterior parietal cortex and in its targets, where those signals can

produce saccades by driving some classes of neurons above a threshold" (Glimcher 2003, 157). These results began to lay the groundwork for understanding more complex decision-making processes, decisions that involve values.

2.1.2 Value-Based Decision-Making

The simple sensorimotor decisions of the previous section mostly involve making motor action selections based on sensory and motor information. However, when animals need to make more complex decisions, plan more complicated actions, or flexibly adapt to changes either in the environment or internally, this involves goal-directed behaviours. Goal-directed behaviour processes involve many more regions of the brain and, importantly, engages the reward network.

The reward network—a complex neural network spanning connections from the cortex, through the midbrain, to the subcortical areas, and back again—involves multiple areas that are of critical importance to the execution of goal-directed behaviour. A "goal" can be understood as an "internal representation of an action outcome currently chosen as the target of the animal's behaviour because of the incentive salience, or motivational value, associated with the outcome" (Mannella et al. 2013, 1). "Value," in this context can be defined as "the subjective costs and benefits that can be attributed to each of the potential outcomes (and associated courses of action) of a decision process" (Gold and Shadlen 2007, 537). We can then think of goal-directed behaviour as "behaviour [that] is sensitive to the value assigned to goals and to the contingencies between goals and actions that can accomplish them" (Mannella et al. 2013, 1).

One of the initial significant insights into value-based decision-making came out of the studies performed by Bechara and colleagues in the nineties (Bechara et al. 1994; 1997). Using brain imaging, Bechara et al. compared the decision-making capabilities of non-lesioned human

subjects, with humans with bilateral damage to the ventromedial prefrontal cortices (vmPFC) (Glimcher 2003). Using the Iowa Gambling Task (IGT), Bechara et al. assessed the risk preferences in the subjects' decisions, using rewards and penalties (Bechara et al. 1994). The study found that normal participants, after a few losses, would adjust their strategies and avoid the card decks with the large losses. However, the subjects with the damage to the ventromedial cortex did not make this adjustment. Instead, the lesioned subjects continuously preferred the riskier decisions, as compared to the normal subjects (Bechara et al. 1994). The results of these studies were an early indicator that the vmPFC must be involved in the storage or updating of value, therefore playing a crucial role in value-based decision-making (Glimcher 2003).

Following these studies by Bechara and colleagues, there was a great deal of research looking into the role of the vmPFC in decision-making. In another card-sorting decision task experiment by Elliott et al. (2000), it was found that in addition to the vmPFC, the midbrain and ventral striatal regions of the brain became active in association with rewards. This study also found that there were activations in the hippocampus that were associated with losses (Elliott et al. 2000). Therefore, these early studies on value-based decision-making were crucial for determining that emotional and motivational information played a large role in decision-making processes.

Since then, the last two decades have seen remarkable progress in identifying the key areas of the basal ganglia—with its connections to the cortex and thalamus, and subcortical areas such as the hippocampus and amygdala—as a crucial area that supports goal-directed behaviour, motivation, and value-based decision-making more generally (Schultz et al. 1997; London et al. 2000; O'Doherty et al. 2000; Rolls 2000; Breiter et al. 2001; Schultz 2006).

It has been suggested by Glimcher (2003) that the basal ganglia (BG), as well as the areas of cortex to which it is functionally connected, may provide nonsensory signals as decision-making inputs to area LIP in the parietal cortex (Glimcher 2003, 163). Gold and Shadlen have similarly claimed that "LIP seems to encode a mixture of sensory, motor, and cognitive signals that might guide decisions about upcoming behavioural responses" (Gold and Shadlen 2007, 364). This would indicate that there might be a shared neural mechanism underlying both sensory and nonsensory decision making, rather than two separate systems in the brain (Glimcher 2003; Sugrue et al. 2005; Gold and Shadlen 2007; Chen et al. 2015). According to Glimcher, the studies reviewed "hint at an extended model of the decision-making system in which frontal and basal ganglia circuits store and update estimates of prior probabilities and reward value, which are used as nonsensory inputs for decision making by the parietal, frontal, and collicular circuits" (Glimcher 2003, 165).

I will return to the idea of a shared neural process for decision-making in the last section of this chapter. However, to consider in more detail the ways in which value-based decision-making processes work and what drives our more complex decisions, we need to first look at the motivation/reward system more carefully. In particular, the next sections consider the role of the basal ganglia in value-based decisions, with its connections to the frontal cortices, thalamus, amygdala, and hippocampus.

2.1.3 Goal-Directed Behaviours and the Basal Ganglia

The basal ganglia (BG) is a group of structures connecting to form pathways (both direct and indirect), integrating information across diverse brain regions, through which emotional, motivational, cognitive, and motor control processing are integrated (Haber 2014). Recent

research focused on the reward network and motivation has suggested that the basal ganglia (BG) structures work together with the frontal cortex and thalamus in order to develop and execute complex, goal-directed behaviours (Haber 2016, 8; see also Schultz, Tremblay, and Hollerman 2000; Frank and Claus 2006; Mannella, Gurney, and Baldassarre 2013; Ikemoto, Yang, and Tan 2015). The BG and cortex work together in order to learn appropriate behavioural responses which importantly, according to Haber, "requires not only the execution of motor plans, but also the behaviours that lead to this execution, including emotions and motivation that drive behaviours, cognition that organizes and plans the general strategy, motor planning and finally the execution of that plan" (Haber 2016, 8).

Researchers have identified the main functional components of the BG, which includes: the ventral striatum (including the nucleus accumbens and olfactory tubercle), the dorsal striatum (including the caudate nucleus and putamen), the globus pallidus, the ventral pallidum, the substantia nigra, and the subthalamic nucleus (Ikemoto et al. 2015). The BG also has strong connections to both the cortex and the thalamus (Haber 2016).

Structurally, the cortico-BG-thalamo-cortical structures are thought to consist of multiple parallel, segregated circuits or "loops" (Mannella et al. 2013). Recent research on the BG and value-based decision-making has primarily been focused on three of these BG circuits, the function of each of these loops is based on the cortical regions that each circuit targets: the limbic loop (motivational), the associative loop (cognitive), and the sensorimotor loop (motor) (Mannella et al. 2013; Ikemoto et al. 2015). The sensorimotor loop includes the dorsal lateral striatum, the premotor cortex, and the primary motor cortex. This circuit is involved in the previously discussed simple sensorimotor decisions, where motor actions are selected based on sensory and motor information (Mannella et al. 2013, 6). This loop is also critical for habitual

behaviours, rather than more flexible, goal-directed behaviours (see Packard and Knowlton 2002; Yin, Knowlton, and Balleine 2004).

The associative loop includes the dorsomedial striatum, and many associative cortical areas such as the inferotemporal cortex, the parietal cortex, and the frontal eye fields (Mannella et al. 2013, 6). This BG circuit is required for orientation, attention, and working memory (Burnod et al. 1999; Hikosaka et al. 2000; O'Reilly and Frank 2006; Cisek 2007) and is involved in both learning and in storing action outcomes (Mannella et al. 2013, 6).

Finally, the third BG circuit is the limbic loop, which includes the nucleus accumbens (or NAc, which is a key part of the ventral striatum), the orbitofrontal cortex (OFC) and the anterior cingulate cortex (ACC). This loop underlies the cognitive processes related to goal-directed behaviour and general executive functions (Mannella et al. 2013, 6).

But how does the limbic loop, crucial for the reward network with its connections between the frontal cortex, basal ganglia, and thalamus, motivate and help us sustain this motivation in order to make us capable of achieving goals? According to a prominent theory of motivation, we can think of both simple and complex goals acting as "rewards" (Kelley and Berridge 2002). These rewards, in turn, help motivate our actions (Hyman 2010, 262). It is clear, then, that the reward system is used in the evaluation of our choices, which motivates our actions. In order to pursue a goal, we must have the capacity to weigh the reward values and make the best decisions in order to achieve these goals or "rewards" (Haber 2011).

Learned environmental cues help us predict the possibility of rewards and become "imbued with incentive properties" (Hyman 2010, 262). These incentive properties, or as it has been termed, "incentive salience," can be understood as the "motivational attribute that the brain assigns to stimuli," as it relates to "the satisfaction of the goals, or rewards, that the animal

requires" (Mannella, Gurney, and Baldassarre 2013; see also Berridge 2004). In turn, these cues activate our feelings of "wanting" and initiate the necessary behaviours needed for obtaining the wanted goal or reward (Bechara et al. 1994; 1997; Hyman 2010, 262). For example, if we were to see someone walking by holding a cup of coffee, just smelling or seeing the coffee might trigger our own "wanting" of a coffee, thereby motivating us to go and buy our own coffee.

More precisely, the reward network is centered on the cortico-basal ganglia circuit (or BG circuit), specifically the limbic loop, which includes connections between the nucleus accumbens (NAc) (part of the ventral striatum, and a key area of the BG), the orbitofrontal cortex (OFC), and the anterior cingulate cortex (ACC). The BG circuit works with the cortex to execute motivated and planned actions (Haber, 2011). The circuits in the reward network respond to and are coordinated through the action of the neurotransmitter dopamine (Montague et al. 2004). Dopamine (DA) is released from the midbrain, in either the ventral tegmental area (VTA) or the substantia nigra (SN). The dopamine is then projected widely throughout the brain, influencing the circuits involved in reward-related learning, cognition, and emotion processes (Haber 2011).

According to reinforcement models, dopamine acts as an error prediction signal, as it is released when a reward is new or better than expected, which aids our learning through the reinforcement of this behaviour (Schultz et al. 1997; Schultz 2006). It is through this BG-DA system that we are able to learn to make choices based on the probability of positive outcomes or rewards (Frank and Claus 2006, 301). Rewards have the ability to guide our behaviours by triggering an increase in dopamine in the NAc (Hyman et al. 2006, 571). Dopamine from the VTA is transmitted to the NAc, as well as to the amygdala and PFC, which also play important roles in the valuation of rewards (Hyman et al. 2006, 571). Overall, the recent research on the

motivation system has shown that we can understand the BG system, with its connections to the PFC and thalamus, as a crucial component underlying value-based decision-making.

2.2 The Orbitofrontal Cortex, the Striatum, and Value-Based Decision-Making

The BG-DA system and its connections to the frontal cortex and thalamus helps to support our ability to learn to make choices and guide our behaviours toward our goals and the rewards that we value. However, we also need to look more carefully at this concept of "value." What is "value," where does it come from, and how does the BG-DA system use it to guide our choices? As previously defined, "value" is understood as the subjective costs and benefits of potential outcomes that occur in a decision-making process (Gold and Shadlen 2007, 537). That is, the focus here is "motivational value," the values that drive our decisions and push us toward what we want. This section will focus on two areas of the brain that are crucial to the valuation process, and therefore central to value-based decision making: the orbitofrontal cortex (OFC) and the striatum—particularly, one part of the ventral striatum, the nucleus accumbens (NAc).

2.2.1 The Orbitofrontal Cortex and Value Representations in Decision-Making

As discussed in the previous section, the BG, aided by dopamine, learns to make choices based on the probability of positive reward outcomes; however the BG needs the orbitofrontal cortex (OFC) to provide active working memory of the expected reward value of a decision (Frank and Claus 2006, 301). The central role for the OFC in decision-making is that it provides internal representations of rewards and works to assign reward values, thereby enabling the necessary calculation of outcome values for decisions. Additionally, the OFC helps in monitoring whether the outcome fulfills our motivational needs and enables the different

alternative decision outcomes to be compared (Strauss et al. 2014, S109). The OFC, located in the frontal cortex, is unique because of its strong connections to all of the sensory areas of the brain (Rudebeck and Murray 2014, 1146). With visual, gustatory, olfactory, and visceral sensory inputs converging in the OFC, the OFC is able to encode a "unified representation of specific outcomes," during the decision-making process (Rudebeck and Murray 2014, 1146).

Initial research on the OFC suggested that the OFC contributed to decision-making by suppressing unwanted behaviours and emotions (Mishkin 1964; Dias et al. 1996; Roberts and Wallis 2000; Rolls 2000). However, more recent studies have shown that this inhibitory role for the OFC does not capture many of its crucial functions in decision-making (Rudebeck and Murray 2014, 1143). Instead, newer studies have shown that the OFC predicts specific outcomes and maintains the updated value representation of these outcomes (Rudebeck and Murray 2014, 1143; Tremblay and Schultz 1999; Elliott et al. 2000; Schoenbaum and Roesch 2005). The OFC encodes a unified representation of specific outcomes held in working memory, which enables goal-directed behaviour by providing information about whether outcomes or action plans need to be updated (Strauss et al. 2014, S109).

Recent studies on the OFC and its connections with the BG have also shown that dopamine neurons in the ventral tegmental area (VTA) use the predicted outcome signals from the OFC in order to compute their reward prediction errors (Takahashi et al. 2011), therefore enabling learning in the brain (Rudebeck and Murray 2014, 1152). Without the input from the OFC, the DA prediction error signals were unable to be updated through experience, which is crucially needed for associative learning (Rudebeck and Murray 2014, 1152).

Further interesting insights have been made about the OFC recently, suggesting a functional distinction between the medial OFC and lateral OFC. The lateral OFC may be

primarily involved with evaluating different outcome options (Rudebeck and Murray 2011; Walton et al. 2010), whereas the medial OFC compares between these options and helps select a goal for action (Rudebeck and Murray 2014, 1150; for studies showing the role of the medial OFC, see Boorman et al. 2009; FitzGerald, Seymour, and Dolan 2009; Noonan et al. 2010; Rudebeck and Murray 2011; Hunt et al. 2012; Strait, Blanchard, and Hayden 2014).

Given all of the information on how the OFC functions in relation to decision-making, how do the signals it produces actually influence our goal selections? First, it is likely that the OFC, through its predicted outcome signals is able to influence and guide action selection in the premotor cortex through its connections to the dorsal lateral prefrontal cortex (dlPFC) (Rudebeck and Murray 2014, 1152; Kennerley and Wallis 2009; Tsujimoto et al. 2011; Cai and Padoa-Schioppa 2014).

Second, the OFC may influence the selection of actions through its direct projections to the striatum, both the ventral and medial striatum regions (Rudebeck and Murray 2014, 1152). This is especially so because of the OFC's influence on DA prediction error signals in the VTA, which sends these signals to the ventral striatum. However, the OFC has dense connections to the medial striatum, as well as the ventral striatum. It is through these direct connections with the striatum that the OFC is said to be able to help regulate habitual versus goal-directed action (Rudebeck and Murray 2014, 1152; see also Gremel and Costa 2013).

This research demonstrates how the OFC uses value representations to guide our decision-making process. However, we still need to consider where the OFC gets its reward information. One key area the OFC receives reward information from is the basal lateral nucleus of amygdala (BLA) (Frank and Claus 2006; see also Holland and Gallagher 2004). A 2002 study by Baxter and Murray showed that the BLA is able to encode both positive (rewarding) and

negative (aversive) outcomes (Frank and Claus 2006, 303; see also Baxter and Murray 2002). Many studies support the idea that the OFC then maintains the reward information that it has received as a value representation in working memory for the purpose of guiding behaviours (Schoenbaum et al. 2009; Hikosaka et al. 2000; Pickens et al. 2003; Rolls 2004; Roesch and Olson 2004). As previously discussed with the studies by Bechara and colleagues (Bechara et al. 1997), patients with damage to the OFC are unable to maintain reward value representations, preventing them from avoiding risky choices and successfully applying goal-directed behaviour more generally.

A study by Winstanley et al. (2004) explored the functional interaction between the OFC and the basolateral amygdala (BLA). They found that, while the OFC and BLA work together to guide goal-directed behaviour, the BLA's primary role in this is to encode the motivational value of a stimulus, whereas the OFC monitors and updates the representations of the predicted outcomes associated with the stimuli. Interestingly though, this study also suggested that the nucleus accumbens (NAc) and BLA may be part of a common circuit that mediates impulsive choices, and maintains a representation of a large, but delayed reward (Winstanley et al. 2004, 4721). However, both the OFC and BLA are considered important nodes in the limbic cortico-striatal circuit, which links them to areas like the nucleus accumbens (NAc) in the ventral striatum, another critically important area in the motivation/reward network (Winstanley et al. 2004, 4721). This brings us to how the striatum, in conjunction with dopamine neurons (DA), plays another key role in value-based decision making.

2.2.2 The Striatum and Dopamine Signaling in Motivation and Reward Processes

The capacity to make value-based decisions and pursue goal-oriented behaviours involves the ability to evaluate and predict risk and rewards, as well as learning from previous experiences in order to avoid poor choices (Haber 2014). In order to do this, the BG must be functioning properly in order to be able to integrate information across motivational, cognitive, and motor control circuits, thereby allowing the development and execution of appropriate action plans (Haber 2014). When properly functioning, "the BG is now recognized to mediate the full range of behaviours leading to the development and execution of action plans including the emotions, motivation, and cognition that drive them" (Haber 2014, 249). Midbrain DA neurons play a key role in reward, cognition, and motor control behaviours, and are associated with both reward-based learning and goal-directed behaviours (Schultz 2002; Haber 2014). However, different subpopulations of DA have been found to be associated with particular functions: the mesolimbic pathway (reward), the mesocortical pathway (cognition), and the nigrostriatal pathway (motor control) (Haber 2014).

From the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc), a massive DA output is sent to the striatum, which is the main input structure of the BG (Haber 2014, 2016). Projections to the striatum come from three main sources: the cortex, thalamus, and the brainstem's dopaminergic cells (Haber 2016, 9). However, the striatum is functionally divided according to the specific cortical input it receives. That is, the ventral striatum (VS) receives input from the OFC, ACC, as well as the amygdala and hippocampus—areas involved in reward, motivation, and emotional processing more generally (Haber 2014, 6). The central striatum receives input from executive and cognitive control cortical areas (Haber 2014, 6). And

finally, the dorsolateral striatum receives inputs from motor control regions of the cortex and is therefore associated with motor control behaviours (Haber 2014, 6).

According to Haber, "the interface between different striatal regions via the midbrain DA cells is organized in an ascending spiral interconnecting different functional regions of the striatum" (Haber 2014, 7). Though the BG circuits are separate, parallel networks, it seems that they do not work in isolation. Through this spiral organization, information can be transmitted between limbic, cognitive, and motor circuits, providing a way in which motivation and cognition can influence motor decision-making processes (Haber 2014, 7). Communication between these functional circuits is crucial in order to be able to adjust our behaviours and make good choices based on new information (Haber 2014, 7). Current work focusing on the striatum has identified "hubs" within the striatum that seem to help coordinate information across the different functional areas (Haber 2014, 7). BG deficits caused by too much or too little DA have been linked to a variety of physical and mental health conditions from Parkinson's Disease and Huntington's Disease, to Obsessive Compulsive Disorder, Tourette's Syndrome, and Schizophrenia (Haber 2014).

DA neurons project from the ventral tegmental area (VTA) to the ventral striatum, whereas DA neurons from the substantia nigra pars compacta (SNc) project to the dorsal striatum (Haber 2016, 9). The ventral and dorsal striatum can be functionally distinguished, with the dorsal striatum playing a key role in behaviours such as habit formation and motor control. The ventral striatum's key function is in reward, goal-directed behaviours, and motivational value. In particular, the nucleus accumbens (NAc) area of the VS has been shown to play a critical role in the encoding of motivational value. According to the literature on the ventral striatum by Haber and colleagues, "convergent fibers from cortex within the VS, taken together with hippocampal

and amygdalo-striatal projections, place the ventral striatum in a key entry port for processing emotional and motivational information that, in turn, drives basal ganglia output" (Haber 2014, 8).

As discussed in the previous section, we know from recent research that midbrain dopamine neurons (DA) play a critical role in motivation and reward; however, there has been some debate over its exact role (Bromberg-Martin et al. 2010). One previously discussed theory is that DA neurons encode motivational value, acting as a prediction error signal and encoding both rewarding and aversive events (Schultz et al. 1997; Berridge and Robinson 1998; Wise 2004). However, a second theory is that DA neurons are encoding motivational salience, signaling and responding to situations or stimuli of high importance (Redgrave et al. 1999; Horvitz 2000; Joseph et al. 2003; Kapur 2003).

In light of these two different roles that DA neurons seem to be playing, it has been recently suggested that rather than thinking of a single DA signal underlying all of the reward and motivation processes, it is more likely that there are distinct, complementary DA signals impacting different circuits in the BG (Saddoris et al. 2015, 11572). According to the study by Saddoris et al. (2015), researchers found that "real-time dopamine release within the NAc (a primary target of midbrain dopamine neurons) strikingly varies between core and shell subregions" (Saddoris et al. 2015, 11572). That is, the NAc can be functionally divided into the NAc core region and the NAc shell region.

Both the NAc core and NAc shell are involved in motivation and reward processes; however there seems to be an interesting functional distinction between the two regions. According to the findings by Saddoris and colleagues (2015), "in the core, DA dynamics are consistent with learning-based theories (such as reward prediction errors) whereas in the shell,

DA is consistent with motivation-based theories (e.g. incentive salience)" (Saddoris et al. 2015, 11572). What Saddoris concludes is that these findings may demonstrate that DA neurons play different and complementary roles, impacting different reward and motivation-based circuits, thereby allowing a more optimal behavioural response to rewards (Saddoris et al. 2015, 11572).

Research over the last decade on reward and motivation has shown a great deal of interest in the role that the NAc plays in motivation. The NAc was initially known as the "limbic-motor interface" (Mogenson et al. 1980). That is, NAc neurons were initially thought to integrate reward-related stimuli with the appropriate motor response in order to pursue that reward (Morrison et al. 2017; see also Mogenson, Jones, and Yim 1980). Newer studies on NAc functions suggest that information about the value of predicted rewards gets encoded by structures that project to the NAc (this includes the amygdala, PFC, and VTA), and is then integrated by the NAc, which in turn is able to influence behaviour (Morrison et al. 2017).

This research is in line with the suggestion by Mannella, Gurney, and Baldassarre (2013), where the NAc may act as the crucial link between values and goals in goal-directed behaviour. Their model proposes that the amygdala and hippocampus are the sources of motivational value, establishing the motivational value of stimuli and goals. The PFC encodes different types of action outcomes—both representing and predicting outcomes, as well as triggering the execution of these actions. Finally, and importantly, the NAc works to integrate motivational value information and selects the goals based on this information (Mannella et al. 2013, 1). They suggest "the NAc integrates different sources of value, representing them in terms of a common currency, with the aid of dopamine, and thereby plays a major role in the selection of action outcomes within the PFC" (Mannella et al. 2013, 1). The goals then pursued are the outcomes that are selected by these processes (Mannella et al. 2013, 1).

2.2.3 Motivational Value Driving Goal-Directed Behaviours: Differences Between "Wanting" and "Liking"

So far, I have considered how motivational value drives value-based decision-making processes. The motivational system—which relies on dopamine signals—underlies our ability to pursue goal-directed behaviours. The motivational network helps us learn and predict rewards and thereby helps us pursue and attain the rewards that we want. The process of wanting to pursue rewards and the pleasure of receiving those rewards often seem to be closely tied together. Indeed, based on an early "dopamine pleasure hypothesis" by Roy Wise (1980), dopamine was initially considered to be our "pleasure" mechanism in the brain (Berridge and Kringelbach 2015, 656). However, this idea was rejected after many studies demonstrated that DA increases do not reliably increase pleasure, nor does the loss of DA necessarily reduce pleasure (Berridge and Kringelbach 2015, 646).

A major contribution to the reward network/motivation research came from Robinson and Berridge (1993), who argued that there was a distinction between what they termed "wanting," the motivational incentive salience controlled by dopamine as discussed in the previous section on the reward network, and "liking," the hedonic reactions, or pleasure felt by consuming rewards[1] (Robinson and Berridge 1993; Berridge 2007; Berridge and Kringelbach 2015; Anselme and Robinson 2016). According to their view, the "liking" process does not involve dopamine, but rather discrete brain regions termed "hedonic hotspots." These hotspots are found in the NAc, VP, OFC, and insular cortices (Castro and Berridge 2014) and act on

[1] Following Robinson and Berridge (1993), the quotations surrounding the terms "wanting" and "liking" are meant to signify that these are unconscious processes, differentiated from conscious feelings of wanting and liking.

opioid and endocannabinoid neurotransmitters rather than dopamine (Anselme and Robinson 2016).

While there is overlap between the "wanting" and "liking" processes, as both come from within common mesolimbic structures, interestingly, research has suggested that they might also be dissociable systems (Robinson et al. 2015). This means that while we might often "want" what we "like," or pursue rewards that give bring us pleasure, we can also "want" without "liking," and "like" without "wanting" (Anselme and Robinson 2016; Robinson et al. 2015). Some examples of "wanting" without "liking" have come from recent research on Parkinson's Disease treatments. With Parkinson's Disease, patients experience motor control issues stemming from the progressive dying off of dopaminergic neurons in the substantia nigra pars compacta (SNc) area of the brain, leading to low levels of dopamine (Meder et al. 2019). Treatments for Parkinson's include dopamine replacement medications, which have been quite effective. However, a small population of Parkinson's patients using these DA agonist treatments have reported the side effect of compulsive motivations, such as gambling, shopping, and other addictive behaviours (Berridge and Kringelbach 2015, 646). While these patients have reported these intense feelings of "wanting" aimed at the addictive behaviours, they have not reported pleasure associated with the behaviours (Berridge and Kringelbach 2015, 646). In other words, they overwhelmingly "want" to do these behaviours, but do not "like" them.

Similarly, research in addictive drugs has also demonstrated the dissociation between "wanting" and "liking." People with substance abuse disorders, especially those with highly addictive drugs like opioids, show intense cravings for drugs while reporting either no increase in pleasure when the drugs are attained or sometimes even a decrease in pleasure experienced when consuming the drug (Anselme and Robinson 2016, 125). The dissociation of these two

systems can explain why an individual with an addiction can both excessively want a drug and not like the drug at the same time.

According to this approach, in addition to the separation of "wanting" and "liking" processes, Anselme and Robinson (2016) have argued that we should understand these processes as nonconscious—distinguishable from our conscious wants and likes (Anselme and Robinson 2016, 125). Their claim that incentive motivation (the "wanting" process) can unconsciously exert control over behavioural performance (Anselme and Robinson 2016, 130) has been particularly well supported by research on addiction. As discussed by Anselme and Robinson (2016),

> Activation of the mesolimbic systems even occurs when drug-related cues are presented outside of conscious awareness (Childress et al. 2008), suggesting that craving and excessive "wanting" for drugs and their cues is the result of mesolimbic dopamine and does not require conscious processing to impact behaviour. (Anselme and Robinson 2016, 126)

The Childress et al. (2008) study also showed that unseen, nonconscious visual triggers were able to activate our motivational "wanting" processes. This activation occurred with both unseen cues for cocaine, as well as with a "natural" or biological reward, such as a sexual cue. Other studies (Fischman and Foltin 1992; Winkielman et al. 2005) also support Anselme and Robinson's conclusion that rewards can unconsciously influence our behaviours (Anselme and Robinson 2016).

Further support for this idea has come from various studies focused on nonconscious "liking" (Winkielman et al. 2005; Pessiglione et al. 2007; Childress et al. 2008; Bargh et al. 2012), where we see that "unconscious liking reactions still effectively change goal-directed

behaviour, though those changes may remain undetected or be misinterpreted even by the person who has them" (Berridge and Kringelbach 2015).

This disconnect between our conscious desires and actual behaviours is regularly seen in everyday life. We may not always be able to accurately say why we did something, or why we like something. Conversely, we may consciously decide to do something that we do not follow through on—like the common New Year's Day resolutions to start exercising and eating healthy that are quickly given up, if started at all. This may suggest that our conscious reasons for deciding on an action may not be terribly effective in terms of motivating and pursuing our goals. This idea underlies the incentive motivation research, which suggests, "conscious goals require first the recruitment of unconscious 'wanting' for cognitive motivations to exist" (Anselme and Robinson 2016, 130). This idea will be further discussed in later chapters involving our subjective experiences of agency and control.

2.3 A Shared Neural Mechanism for Decision-Making: Integration of Cognitive, Emotional, and Motivational Information

The decision-making research examined in this chapter, demonstrates a clear shift away from understanding decision-making as a cortical-based, highly cognitive process that is distinct from emotional and motivational processes. Instead, the research suggests an understanding of decision-making as an integrated system. This integrated decision-making system, encompassing widely distributed areas of the brain (both cortical and subcortical), involves cognition, emotions, and motivational processes supported by the reward network.

Although this chapter is divided into separate sections dealing with perceptual decision-making and value-based decision-making separately in order to show the historical development

of decision-making research, emerging evidence suggests that there may not be such a strict division between the two types of decision-making processes (Sugrue et al. 2005; Gold and Shadlen 2007; Mannella et al. 2013). As discussed in Sugrue, Corrado, and Newsome (2005),

> Midbrain dopamine neurons broadcast their error signal to both the striatum and the cortex, and as the functional scope of that signal has yet to be defined, biologically it seems more appropriate to view such decision-making systems not as distinct entities, but as the theoretical extremes of a continuum of processes that are responsible for valuation and action. (Sugrue et al. 2005, 368)

As suggested by Sugrue and colleagues (2005), perceptual- and value-based decision-making should be considered as two ends of a continuum of decision-making processes. This was also recently argued by a research group trying to bring together the two types of decision-making (Bogacz and Gurney 2007; Bogacz and Larsen 2011; Lepora and Gurney 2012; Mannella et al. 2013). According to the summary of this research by Mannella, Gurney, and Baldassarre (2013), "the idea is that evidence for a decision about a particular percept is accumulated as cortical activity in high level sensory areas and that is used as input into BG working as a selection engine for action" (Mannella et al. 2013, 22). A review of decision-making by Sugrue, Corrado, and Newsome (2005) has similarly claimed that the study of value-based decision-making needs to bring together neuroscience research from historically separate areas of research: perception and cognition, and motivation, and reward (Sugrue et al. 2005, 374).

A recent article by Chen et al. (2015) has also shown that reward can influence perceptual decision-making. Interestingly, they suggest that value can bias perceptual decision-making "through modulating multiple systems (perceptual-, decision-, and action-networks) instead of a single top-down process, as usually suggested in the literature" (Chen et al. 2015, 12). Similar

research has been done by Pessoa (2013), suggesting that the amygdala may contribute to the biasing of value representations (as well as contributing to positive value representations) in decision-making (Pessoa 2013, 244). According to Glimcher, recent research on decision-making suggests a system of frontal and BG circuits storing and updating reward values, "which are used as nonsensory inputs for decision making by the parietal, frontal, and collicular circuits" (Glimcher 2003, 165).

The idea of a shared mechanism, or at least decision-making processes that involve deeply integrated cognitive, emotional, and motivational information is consistently suggested throughout the research, occurring across many groups working on diverse aspects of decision-making and decision-making related processes. In particular, the literature surrounding the BG and its function in decision-making, demonstrates the importance of the integration of motivational, emotional, and cognitive information during decision-making.

Conclusion

Decades of research on the neural processes that underlie our decision-making processes have begun to show the complexity of just how many processes and widely diverse areas of the brain are needed for even simple decisions—never mind the complex decisions and goal-directed behaviours that we take on every day. This chapter has focused on reviewing the pertinent research on these processes in order to investigate the idea of what motivates us to act, to decide on certain choices, and to pursue what we want in life. How do we make everyday decisions, and strategically plan long-term actions? These are all questions that need to be determined in order to get a more complete understanding of what it means to make autonomous decisions.

As reviewed, recent research focused on value-based decision-making and motivation has suggested that the basal ganglia (BG) structures work together with the frontal cortex and thalamus, via the cortico-striatal-thalamo pathway, in order to develop and execute complex, goal-directed behaviours (Haber 2016). The BG-DA system and its connections to the frontal cortex, thalamus, as well as the amygdala and hippocampus are crucial to our ability to learn to make choices and guide our behaviours toward our goals and the rewards that we value.

Also reviewed in this chapter was research relating to motivational value and the valuation process. This section focused on both the orbitofrontal cortex and striatum (particularly the NAc area of the ventral striatum) as having critical roles in constituting value in decision-making and goal-directed behaviour, as well as in integrating cognitive, emotional, and motivational information across various brain regions.

The BG structures learn to make choices based on the probability of positive reward outcomes (with the aid of dopamine neurons); however, the BG importantly needs the orbitofrontal cortex (OFC) to provide active working memory of the expected reward value of a decision (Frank and Claus 2006, 301). The OFC provides the internal representations of rewards and assigns reward values, enabling the calculation of outcome values for decisions. Studies reviewed on the role of the striatum suggest that information about the value of predicted rewards gets encoded by structures that project to the NAc, such as the PFC, VTA, and amygdala, which in turn gets integrated by the nucleus accumbens (NAc).

Also discussed in this chapter was the fact that motivational "wanting" and "liking" might be dissociable systems (Anselme and Robinson 2016). At times these two systems can be closely tied together, meaning we might actually "want" what we "like," or pursue rewards that give bring us pleasure. However, this research has also emphasized the interesting point that we

can also “want” without “liking,” and “like” without “wanting” (Anselme and Robinson 2016). This line of research has been shown to have interesting implications for the realm of addiction, in particular.

Importantly, this chapter has demonstrated the shift in thinking about decision-making as a strictly cortical process, distinct from emotional and motivational information in subcortical areas of the brain. Instead, the most current research suggests an integrated decision-making system, encompassing widely distributed areas of the brain (both cortical and subcortical); that is, understanding decision-making as a process that necessarily involves cognition, emotions, and motivational processes.

Chapter 3

The Neuroscience of Control: Goals, Emotion, and Motivation in Control Processes

Introduction

Like the integrated cognitive-emotional-motivational accounts of decision-making just discussed, there is emerging research on the role that emotional and motivational information plays in the neuroscience of self-control as well. While discussions of self-control in the philosophical autonomy literature are common, it is often portrayed as a highly rational process where we consciously and effortfully try to resist affective temptations. However, I argue that this portrayal of self-control needs to be updated to reflect a stronger role for affect in the control process itself. This chapter aims to examine the concept of self-control more closely by looking at the recent neuroscientific research on this subject, which importantly, suggests that emotional and motivational information is crucial to the process of self-control.

The first section concentrates on distinguishing between three related control processes: self-regulation, self-control, and cognitive control. These terms have at times been used interchangeably in the self-control literature, and it is important to clearly distinguish them from one another as they refer to different neural processes. Within this section, I will also explain how goals, conflict, motivation, and emotion impact the three types of control discussed.

Self-control is commonly understood as a process for resolving conflicts between competing goals that are mutually exclusive (Inzlicht et al. 2021, 320). Fundamentally, then, to understand self-control requires investigating the way in which we represent goals and select between conflicting goals. The way in which goal conflicts are framed by self-control models

has important implications for how we come to define self-control, as well as how emotion, motivation, and goals are believed to intersect. Many traditional accounts of self-control stem from dual-process models, where the central claim is that there are two competing systems in the brain, the "cold," deliberative system that represents long-term goals, and the "hot," impulsive system that reacts to immediate rewards. This has contributed to a traditional self-control framework that views reason and emotion as competing systems. We can see this duality in paradigmatic self-control dilemmas that present self-control as a moral choice between maintaining one's long-term goal in the face of an immediate and very tempting short-term reward. I will explore this idea in the last section, where I present alternative frameworks that broaden this narrow conception of self-control. As opposed to traditional models that suggest emotion undermines self-control, I will discuss newer models of self-control that view emotions as a central component of self-control processes.

Overall, in this chapter I argue that the role of affect in the empirical research on self-control has been generally understated. By reviewing the neuroscientific literature of self-regulation, self-control, and cognitive control, we can see that emotion and motivation plays a central role in all three types of control. Furthermore, I also argue that emotion has been mischaracterized in self-control dilemmas, in particular by dual-process models, as undermining our self-control. I argue that emotion does not always undermine self-control, and in fact, it is an essential component in self-control, self-regulation, and cognitive control.

3.1 The Neuroscience of Self-Control

Despite decades of empirical research on self-control and the progress that has been achieved in this area, a simple definition of self-control is not yet established. There is not a

definitive consensus on what self-control is as a concept, nor is there a clear answer of how this process works in terms of its neural mechanisms. Potentially adding to this confusion within the empirical self-control literature is that the terms self-regulation, self-control, and cognitive control are at times used interchangeably when referring to processes that involve control over our goal-directed behaviours (see Inzlicht et al. 2021, 321 for a detailed discussion of this problem).

However, some recent work (Nigg 2017; Inzlicht et al. 2021) has attempted to establish the key differences in various control processes and clarify these concepts. The growing consensus seems to be that while self-regulation, self-control, and cognitive control can at times overlap, they are in fact three distinct processes and refer to different elements of control (Milyavskaya et al. 2019, 80; Inzlicht et al. 2021, 321). This recent research has provided many new insights, clarifying the distinctions between the three concepts. Hopefully this can lead us to more clearly understand each process and the underlying neural mechanisms that support them in the near future. The following three subsections will provide an overview of self-regulation, self-control, and cognitive control processes, as well as a review of the prominent cognitive, psychological, and behavioural neuroscience research models that have provided us with some insight into the underlying neural mechanisms of control.

3.1.1 Self-Regulation

A recent review on the subject defines self-regulation as "the dynamic process of determining a desired end state and taking action to move toward it while monitoring progress along the way" (Inzlicht et al. 2021, 320). The "desired end-state" is understood as a goal (Inzlicht et al. 2021, 321). This definition suggests that we can understand self-regulation as a

broad umbrella term that encompasses many different aspects of goal-directed behaviour—including how we choose which goals to pursue, the monitoring, planning, and implementation of those goals, and shielding the goals from competing actions or thoughts—all of which help a person achieve their desired end-states (Inzlicht et al. 2021, 320). The way a desired-end state or goal is defined within much of the recent literature, does not just refer to a desired behaviour or action, but also refers to desired thoughts and emotions (Inzlicht et al. 2021, 320). "Self-regulation includes the various ways in which people modify their thoughts, feelings, and behaviours in the service of a personal goal…" (Inzlicht et al. 2021, 320). Therefore, self-regulation as a concept is quite broad, as it goes beyond just the regulation and modification of our actions, including all of our thoughts and emotions as well (Inzlicht et al. 2021, 320).

With self-regulation being such a broad concept, referring to many different underlying processes, it no surprise that there are many different networks supporting these goal-relevant activities. Within the self-regulation literature, the neural mechanisms are often broken down into two main sub-groups: top-down and bottom-up self-regulation processes. The two types of processes are to be understood as functioning on a continuum rather than two distinct categories, due to the continual reciprocal neural feedback loops (Nigg 2017, 365). That is, they are not separate competing processes, as would be argued from a dual-process account. Rather, top-down and bottom-up regulation refers to the different ways in which self-regulation can be activated and regulate our responses. Furthermore, both top-down and bottom-up approaches are thought to involve widely distributed circuits across the brain, involving both cortical and subcortical areas. For example, as mentioned in another recent literature review on self-regulation, "top-down systems can activate, suppress, or bias bottom-up responses. In turn, bottom-up systems can activate goal-related behaviour via priming or other effects, limiting the

effects of top-down processes" (Nigg 2017, 365). Therefore, to understand how self-regulation works, we should consider each of the sub-groups of self-regulation processes, and how they impact each other to produce and direct our goal-directed behaviours.

Bottom-up self-regulation processes often refer to processes that are automatic, rapid, and largely driven by the valuation and reward systems. As described by Nigg (2017), "they are called "bottom-up" as shorthand because they are elicited by sensory (external) stimuli and because human brain imaging and primate single cell recording studies link them with 'feed-forward' neural signaling (e.g. subcortical to cortical, or posterior to anterior cortical signaling)" (Nigg 2017, 365).

While these bottom-up processes are usually discussed as the targets of regulation (e.g., when someone needs to regulate their emotional responses), Nigg (2017) lists at least four ways bottom-up regulation processes are also regulatory themselves (Nigg 2017, 365). The first way they can be considered regulatory is that bottom-up processes, such as incentive approach and avoidance, modulate each other through associative and instrumental learning history (Nigg 2017, 365).

Second, bottom-up processes can often be seen to "prime, activate, or modulate top-down processing" (Nigg 2017, 365). Approach and avoidance reactive processes are considered an important part of self-regulation of action and cognition (Nigg 2017, 365), and play a crucial role in reorienting attention and inhibiting behaviour in situations of uncertainty, allowing us to quickly adapt to novel, uncertain, or potentially harmful situations (Nigg 2017, 365).

Third, through learning processes, top-down activities can be automatized in particular contexts—allowing for bottom-up automatic responses that continue to be regulatory with very little effort. As discussed later on in the cognitive control literature, it is theorized that we

automatize as much as possible to avoid the costs associated with exerting cognitive control processes (Nigg 2017, 365).

Fourth, and last, bottom-up processes are regulatory in the sense that they provide us with important goal-related information, such as threat or reward information, or previously learned associations when cued (Nigg 2017, 365). The bottom-up processes involved in self-regulation play a valuable role in regulating our behaviours and keeping our goal-directed behaviour on-track, particularly through its connections to our motivation and value systems.

On the other hand, top-down self-regulation processes have more traditionally been the focus of research and discussion in relation to self-regulation. Top-down processes of self-regulation are thought to activate when novel problems or response conflicts arise (Nigg 2017, 366). According to Nigg (2017),

> They are called top-down because they respond to internal mental representations (such as a goal or a rule) rather than external (sensory) stimuli, and because human imaging and primate single cell recording data link them to 'feed backward' neural signaling (i.e. cortical to subcortical or anterior to posterior cortical). (Nigg 2017, 366)

There has been some difficulty in identifying the neural processes specifically involved in self-regulation processes because there is so much overlap in processes that support self-regulation, as well as other executive functions in the brain (Nigg 2017). However, there is a consensus on a few particular nodes playing a central role including the anterior cingulate cortex (ACC), ventromedial prefrontal cortex (vmPFC), ventrolateral prefrontal cortex (vlPFC), and the dorsolateral prefrontal cortex (dlPFC) (Nigg 2017, 366). As described by Nigg (2017), "these nodes are embedded in circuits (such as the ACC-DLPFC circuit). The circuits in turn are embedded within distributed functional and anatomical cortical-subcortical networks" (Nigg

2017, 366). Nigg further explains that two important networks in top-down regulation are the cingulo-opercular and fronto-parietal networks:

> The cingulo-opercular network, broadly, includes the inferior regions in lateral PFC (including ventrolateral PFC), regions of the ACC, the insula, and other nodes. The fronto-parietal network, broadly, involves superior regions in lateral PFC, including dorsolateral PFC, as well as regions of posterior parietal and inferior temporal cortex and subcortical connections. Cross-talk with additional brain networks is important, including the default mode network (which includes the ventromedial PFC) and a noradrenergic, locus-coeruleus 'alertness' network involved in arousal regulation. Subcortical regions, particularly the thalamus and hippocampus, also participate indirectly in top-down regulation via information salience and updating. (Nigg 2017, 366)

According to Nigg's (2017) review, when we use the term self-regulation, both top-down and bottom-up processes should be considered in terms of describing the underlying processes. However, depending on which aspect of self-regulation (as a broad, umbrella term) is the particular topic of discussion, the specific brain regions or networks that are involved in the underlying processes may differ accordingly. For example, the activity of monitoring goals will have different mechanisms than the activity of emotion regulation (Nigg 2017).

Relatedly, there are also differences in the underlying mechanisms when we look at typical self-regulation failures in the literature. Self-regulation failures are when things go awry in adaptive goal-directed behaviours and include issues with disinhibition, impulsivity, and risk-taking behaviours (Nigg 2017, 373). These terms are sometimes taken to mean the same kind of break-down of self-regulation and often used interchangeably. However, there are important differences, as specified by the Nigg (2017) review. These differences are particularly relevant in the autonomy literature, as examples of failures of autonomy often include cases of impulsivity, disinhibition, and inappropriate risk-taking behaviours.

Response inhibition pertains to the interruption of a response, or in other words, it can be thought of as the brain's "brake." This function is needed to stop, pause, and change responses (Aron et al. 2004). Response inhibition failures are often seen to be synonymous with impulsive behaviours. However, while response inhibition can be a component of impulsivity, they are not interchangeable concepts. In terms of the underlying neural mechanisms of response inhibition, it "largely relies on a brain network involving mainly right lateralized frontal brain areas, such as inferior frontal gyrus (IFG) and pre-supplementary motor area (pre-SMA), projecting to the motor cortex via the cortico-striatal-thalamic cortical (CSTC) connections" (Gillan et al. 2016, 832).

Within the literature on inhibition, the Inferior Frontal Gyrus (IFG) has been consistently associated with self-inhibitory processes, including motor, cognitive, and affective inhibition (Aron 2007; Tabibnia et al. 2011; Aron et al. 2014; Bari and Robbins 2013). It has been suggested that it is part of a larger inhibitory network through its direct and indirect connections to the pre-supplementary motor area (pre-SMA), basal ganglia (BG), and importantly the subthalamic nucleus (STN), which inhibits the response of the basal ganglia (Bari and Robbins 2013). The cortico-basal ganglia connections that project to the STN are part of what has been termed the "hyperdirect pathway." As explained by Bari and Robbins (2013), this pathway "allow[s] for fast inhibition of ongoing actions by increasing inhibitory signals from the globus pallidus, to which the STN strongly projects, thus inhibiting the output from the basal ganglia" (Bari and Robbins 2013, 62).

Another area associated with stopping tasks is the pre-supplementary motor area (pre-SMA), which is functionally and structurally connected to the right inferior frontal cortex (rIFC) (Aron et al. 2014, 177). While further research is needed to establish whether the "rIFC triggers

the STN directly, or via the pre-SMA…the data show that the rIFC is embedded in a three-way fronto-basal-ganglia network for outright stopping…that implements inhibition via downstream effects on premotor and primary motor cortex" (Aron et al. 2014, 178).

Most of the research concerning motor and affective control has come from neuroimaging studies showing inferior frontal gyrus pars opercularis (IFGpo) activity (a subsection of the IFG) in healthy individuals and from studies of disorders involving inhibitory control and IFG deficits (Tabibnia et al. 2011, 4806). Deficits in inhibitory control and in the IFG are seen in many different psychiatric disorders, such as depression, attention deficit hyperactivity disorder (ADHD), obsessive compulsive disorder (OCD), and various addictions (Tabibnia et al. 2011, 4805). For example, individuals with addictions to methamphetamines have been shown to have poor motor inhibitory control and emotion dysregulation, as well as structural and metabolic abnormalities in many areas of the brain, including the IFG (Tabibnia et al. 2011, 4805).

The role of the IFG in emotion regulation has been shown in a few different studies, demonstrating specifically that the down-regulation of negative emotions was associated with IFG activity (Hariri, Bookheimer, and Mazziotta 2000; Ochsner et al. 2004; Phan et al. 2005; Lieberman et al. 2007; see Tabibnia et al. 2011, 4808). Relatedly, a recent study by Volkow et al. (2010), has shown that the right Inferior Frontal Cortex (rIFC) seems to be involved in the control of cravings as well. This study showed that when individuals with a drug addiction purposefully tried to inhibit their cravings, there were decreases in activity in reward and reward-related areas such as the nucleus accumbens (NAc) and medial orbitofrontal cortex (mOFC). The decreases in reward related activity in the NAc and mOFC were correlated with increases in activity in the rIFC (Volkow et al. 2010, 2538).

The recent research on inhibition has led to the proposal that the "right inferior frontal cortex, in particular, the right inferior frontal gyrus (rIFG) (along with one or more fronto-basal-ganglia networks) is best characterized as a brake" (Aron et al. 2014, 177). It has been suggested that "this brake can be turned on in different modes (totally, to outright suppress a response; or partially, to pause) and in different contexts (externally, by stop or salient signals; or internally, by goals)" (Aron et al. 2014, 177). As it stands, we can consider the right IFG region as crucially involved in both affective and non-affective inhibitory control, implementing this inhibitory control through its connections to a wider prefrontal-basal ganglia network (Aron et al. 2014, 183).

While response inhibition and impulsivity can overlap in some situations, as previously mentioned, impulsivity is not synonymous with disinhibition. Impulsivity can include certain behaviours like inappropriate responses to a stimulus, which overlaps with disinhibition. However, it also involves processes related to the weighting of rewards, both immediate and delayed (Nigg 2017, 373). The processes with which we calculate the value of rewards and the temporal component of delaying rewards are not included in inhibition processes and is mostly modulated by bottom-up processes related to reward valuation (Nigg 2017, 373).

A recent article on impulsivity (Dalley and Robbins 2017) has suggested that there are actually different types of impulsivity. Each of the types of impulsivity are related to distinct psychological and neural mechanisms (Dalley and Robbins 2017, 158). As the authors explain, there can be quite a bit of variation in behaviours that we might call impulsive, where "impulsive behavior can be related to both increased motivation and reduced motivation (apathy) and it can represent either a failure to process information sufficiently or to control response output"

(Dalley and Robbins 2017, 158). Daley and Robbins break down impulsivity into two main groups: decisional impulsivity and motoric forms of impulsivity (Dalley and Robbins 2017, 159).

Within the decisional impulsivity group, there are three subgroups: temporal discounting, probalistic discounting, and reflection impulsivity. Temporal discounting refers to the preference for immediate rewards that are smaller, versus a larger delayed reward (Dalley and Robbins 2017, 159). Probalistic discounting is related to how risky of a decision you are willing to make and whether there is a greater preference for smaller, more likely rewards versus larger, less likely rewards (Dalley and Robbins 2017, 159). And last, reflection impulsivity refers to a tendency to make quick decisions without waiting to acquire adequate information, or without considering all the evidence that is available (Dalley and Robbins 2017, 159).

Within the second, motoric impulsivity group, there is also a further break down into two main subgroups: stopping impulsivity and waiting impulsivity (Dalley and Robbins 2017, 159). These two groups are differentiated by particular tests, the stop-signal reaction time (SSRT) task and the premature responding task (Dalley and Robbins 2017, 159). The SSRT task measures one's ability to stop a response after it has already been initiated and is referred to as stopping impulsivity (Dalley and Robbins 2017, 159). Waiting impulsivity, on the other hand, is assessed with premature responding tests that assess the ability to resist responding before a certain period of time has gone by (Dalley and Robbins 2017, 159).

In terms of the neural networks that mediate these different types of impulsivity behaviours, it has been suggested that interactions between the striatum, prefrontal cortex, and hippocampus are central, along with neuromodulation of the ascending monoamine systems (which includes dopamine (DA), serotonin (5-HT), noradrenaline (NA), and histamine (HA) circuitries (Dalley and Robbins 2017, 159). In particular, the nucleus accumbens (NAc), which

as discussed in the previous chapter receives dopamine (DA) input from the ventral tegmental area (VTA), is considered a central structure for impulsivity (Dalley and Robbins 2017, 159).

However, the ways in which these networks work differ according to the type of impulsivity discussed. For example, Dalley and Robbins (2017, 159) discuss the differences in NAc and dopamine involvement in premature responding tasks versus temporal discounting tasks (i.e., waiting impulsivity, a form of motoric impulsivity, versus temporal discounting, a form of decisional impulsivity):

> The two distinct task-related expressions of impulsivity described to be affected by changes to the NAc (that is, failure to supress premature responses, and failure to delay gratification) may depend on subtly different mechanisms within this structure. Whereas the capacity to delay gratification is associated with decreased DA release in the NAc core, impulsive premature responding is associated with decreased DA release in the core and increased DA release in the shell subregion… in general, it seems that premature responding and temporal discounting may be mediated primarily by distinct regions of the NAc. (Dalley and Robbins 2017, 161)

Therefore, the term impulsivity is more accurately an umbrella term for various types of impulsivities, each represented by slightly different underlying neural mechanisms. It would be more accurate to specify the distinct type of impulsivity, rather than grouping them together.

The last self-regulation failure typically discussed in the self-regulation literature is risk-taking behaviours (Nigg 2017). We can see an overlap of this behaviour within the subgroup of decisional impulsivity—where the weighting of the probability of rewards is the central feature of the choice amounting to a kind of risk-taking impulsivity. However, risk-taking behaviour does not have to be impulsive. For instance, an example used by Nigg is a long-planned trip to climb Mount Everest—a trip with a high chance of injury or even death—is high in risk-taking, but not impulsive if well-planned (Nigg 2017, 374). What makes a high risk-taking decision a

self-regulation failure would be related to whether the rewards are too highly valued, or the possible harms undervalued (Nigg 2017, 373).

In summary, self-regulation is a broad term that encompasses many different activities, which are supported by distinct neural mechanisms. While all self-regulation activities and the underlying processes attempt to serve adaptation and the achievement of goals (Nigg 2017, 375), there are various ways that we go about achieving this ideal end-state. As noted, there is also quite a bit of overlap between the different types of activities that support self-regulation, as well as overlap in the different ways self-regulation can fail. However, a discussion of conceptual distinctions of the many different types of self-regulation can help clarify what is included when referring to this concept, as well as providing a more precise answer about what is happening when these processes break down. At times, self-regulation may include self-control processes, but the two terms are not equivalent despite being used interchangeably. The next section will examine the process of self-control, and which neural mechanisms may be at work supporting this process.

3.1.2 Self-Control

The definition of self-control is still currently debated within the empirical literature. According to a recent review on the subject (Inzlicht et al. 2021), self-control is differentiated from self-regulation based on whether there is a conflict present between two possible goals. That is, self-regulation is an umbrella term referring to any aspect of goal-directed processes, whereas self-control refers to the specific "process of advancing one goal over a second goal when the two come into conflict" (Inzlicht et al. 2021, 321). As described in more detail by Inzlicht et al. (2021),

> Self-control refers to all means of resolving conflicts between competing goals (including, but not limited to, inhibition), whereas self-regulation can be conflict-free, refers to the broader process of steering one's behaviour toward a desired end state, and includes setting goals, monitoring goal progress, and acting in accordance with goals. (Inzlicht et al. 2021, 322)

There are many different models for self-control, various definitions of what it means to be presented with a goal conflict, and what it means to resolve this conflict[1]. This section will review some of the most prominent models of self-control and ways in which self-control has traditionally been defined. Finally, I will discuss some of the newer models and how the definition of self-control has been expanded in recent years. While there may still not be a clear consensus on what self-control is, as well as what specific processes are involved in it, the research is overall largely focused on self-control being a process of resolving goal conflicts (Inzlicht et al. 2021).

Initially, self-control referred to a very specific process of consciously maintaining and pursuing our long-term goals in the face of contrary temptations or impulses. In other words, traditionally, self-control definitions were related to what we think of as "will power." For example, early accounts of self-control (Mischel 1974; Mischel, Cantor, and Feldman 1996; Baumeister et al. 1998) argued that self-control was a way that we consciously and deliberately override impulses and temptations (Gillebaart and de Ridder 2015, 89). According to Gillebaart and De Ridder (2015), this traditional type of effortful self-control presupposes that an "active self" is required to consciously control one's own behaviour, thereby excluding any sort of automatic self-control processes (Gillebaart and de Ridder 2015, 90). Mischel's famous marshmallow experiment is a paradigmatic example of the ongoing conscious effort of

[1] See Inzlicht et al. 2021 for a detailed review of the various self-control models.

attempting to delay gratification. In the marshmallow experiment, children are presented with the choice between an immediate reward of one marshmallow or waiting a set period of time and receiving two marshmallows as a reward for waiting (Mischel 1974). According to this view, self-control consists of the ability to resist the small immediate reward and wait for the delayed, yet larger reward.

A prominent account of self-control focused on this type of temporally asymmetric goal conflict is the resource model of self-control (Baumeister et al. 1998; Baumeister, Tice, and Vohs 2018). On this view, self-control is also presented as a cognitively effortful process of overriding temptations and impulses, which can include thoughts, feelings, and behaviours (Inzlicht et al. 2021, 321). In particular, this model argues that our ability to effortfully override these impulses wanes over time (e.g., the ego-depletion effect) (Inzlicht et al. 2021, 330). In their 2021 review of self-control models, Inzlicht et al. state that there is not much evidence to support the resource model or the ego-depletion effect in self-control processes, despite its decades-long popularity (Inzlicht et al. 2021).

Another highly influential account that defines self-control by the conflict between long-term goals and short-term temptations are the dual-process models of self-control (Cohen 2017; Heatherton and Wagner 2011; Hofmann, Friese, and Strack 2009; Kahneman 2011; Metcalfe and Mischel 1999; Thaler and Shefrin 1981). Dual-process models have been popular in many different domains of research: decision-making, self-regulation, cognitive control, and here in the self-control research as well. Dual-process models argue that there are two distinct systems that guide our behaviours – the “hot” system 1 and “cold” system 2.

On this view, system 1 is the automatic and impulsive system that quickly responds to the environment, and in particular, responds to stimuli that hold a lot of incentive value (e.g., highly tempting rewards) (Inzlicht et al. 2021, 324). According to Inzlicht et al. (2021),

> This system responds to stimuli that are temporally and spatially near, serves short-term gratification, and produces behaviours that are rigid and habitual—most often an urge to approach and act on the temptation at hand (Hofmann et al. 2009). System I devalues stimuli that are temporally distant (Ainslie 1974; Mischel et al. 1972) or are contingent on effort (Apps et al. 2015; Westbrook et al. 2013). This impulsive system is associated with activity in subcortical regions of the brain involved in rewards and emotion, such as the nucleus accumbens in the ventral striatum, the amygdala, and the insula. (Heatherton and Wagner 2011; Lopez et al. 2014) (Inzlicht et al. 2021, 324)

In contrast, system 2 is the cold, deliberative, and reflective system. This system is thought to underlie higher-order mental operations such as setting goals and the planning the strategies to pursue those goals, making careful deliberative judgments and evaluations, and generally is thought to be the 'planner' (Inzlicht et al. 2021, 324). As discussed by Inzlicht et al. (2021),

> Unlike System I, this system is influenced by long-term goals to flexibly respond to the environment, including overriding behavioural tendencies from the impulsive system. Also unlike System I, the control system processes stimuli sequentially, meaning that it is limited by a person's attentional capacity, which can fluctuate from moment to moment based on concurrent processing demands (Cowan 2001). This control system is associated with regions of the lateral prefrontal cortex (Berkman et al. 2011; Heatherton and Wagner 2011). (Inzlicht et al. 2021, 324)

In terms of how these dual-process models look at the self-control process, it is thought that when there is no conflict, generally system 1 is dominant. However, any conflict triggers system 2 to override temptations through conscious deliberation to adapt our behaviour, keeping it on track with our goals. According to this view, system 2 only activates when necessary, in the

face of conflict because of how effortful and costly its functions are. Thus, in order to preserve the resources of system 2, much of our behaviour is guided by the more automatic system 1 (Inzlicht et al. 2021, 330). In a self-control dilemma, when these two systems are in conflict (e.g., when a person on a diet is tempted by a piece of chocolate cake), dual-process models of self-control state that the two systems would compete and whichever more strongly activates, would win—thereby deciding which response is the behavioural outcome (Inzlicht et al. 2021, 325). Therefore, self-control, on this view, can be understood as a battle between cold deliberative and hot impulsive processes which compete for dominance over our behaviours (Inzlicht et al. 2021, 325).

Despite the popularity of dual-process models, more recent models have been moving away from framing self-control as a conflict between long-term goals on the one hand, and impulses, or temptations that conflict with those goals (Gillebaart and de Ridder 2015; Inzlicht et al. 2021). As explained by Inzlicht et al. (2021),

> More recently, the concept of self-control has expanded to include any means to advance one motive over another (Fujita 2011), including means that do not involve effort (Galla and Duckworth 2015; Gillebaart and de Ridder 2015; Hennecke et al. 2019), means that involve inhibiting something undesired or enacting something desired (de Ridder et al. 2011; Gillebaart 2018), or means that involve reactive inhibition or more proactive forms of resolving conflict (Duckworth et al. 2016; Fujita 2011). (Inzlicht et al. 2021, 321)

The newer models of self-control, therefore, include different types of goal conflicts that go beyond the short-term reward versus long-term goal conflict. Furthermore, the way that self-control has been reframed and expanded includes other ways to resolve goal conflicts that go beyond the effortful, conscious inhibition of temptations. These strategies may also be more proactive, rather than inhibiting temptations after the fact. The reframing of self-control in this way also moves us further away from the traditional idea that self-control necessarily involves a

conscious, “active self,” operating as enforcer of our self-control processes (Inzlicht et al. 2021). According to the expanded approaches to self-control, a conscious, active self is not always needed—especially given how much of our goal-directed behaviours operate unconsciously. This new reframing of self-control and the various strategies for dealing with goal-conflicts that go beyond conscious, effortful inhibition will be discussed in more detail when reviewing the alternate models of self-control below.

Turning now to the alternate models of self-control not centrally focused on goal conflict, some examples include cybernetic models of self-control (Carver and Scheier 1998), goal-systems theory (Kruglanski et al. 2002), and choice models of self-control (Berkman et al. 2017; Berkman 2018). The cybernetic model of self-control (Carver and Scheier 1998) has been highly influential and is primarily centered around a target goal based on feedback loops (Inzlicht et al. 2021, 322). Behaviours or responses are assessed in relation to the desired goal, with the system monitoring for conflict or discrepancies between the response and the goal. As described by Inzlicht et al. (2021),

> The output of the implementing system feeds back to change the current state of the organism, after which the current state is again compared to the goal state by the monitoring system. This process is repeated until the discrepancy between the goal state and the current state is reduced to some acceptable level. (Inzlicht et al. 2021, 322)

The conflict is the discrepancy between the current state and the desired end state when the goal is achieved; however, achieving the desired end-state is the primary focus of this model, not the conflict itself (Inzlicht et al. 2021, 330). This model also importantly specifies that the discrepancy (or prediction error) causes negative emotions to arise, whereas the reduction of the discrepancy produces positive emotions (Inzlicht et al. 2021, 330).

Another model that focuses less on conflict, is the goal-systems theory of self-control (Kruglanski et al. 2002). Specifically, this model focuses on the structural organization of goals and their associated means (i.e., the different ways of attaining our goals) (Inzlicht et al. 2021, 323). This model argues that goals and their associated means make up a goal system, which is a hierarchical and interconnected network (Inzlicht et al. 2021, 323). This goal system drives behaviour as the more strongly associated a means is to a goal, the more likely it is to be selected and acted upon (Inzlicht et al. 2021, 323). Interestingly, this approach can be understood to be arguing that goal-directed behaviour is impacted jointly by both cognitive and motivational factors (Kruglanski et al. 2002). As Kruglanski and colleagues argue,

> To that end we defined the concept of goal-systems as a mental representation of cognitively interwoven constructs whose contents consist of "goals" and "means." We have assumed that the functioning of goal-systems is guided jointly by two classes of properties, cognitive and motivational. (Kruglanski et al. 2002, 375)

Related to this notion of goals and their associated means, is an interesting discussion of studies by Kruglanski and colleagues on self-control dilemmas that focus on long-term goals versus temptations (Fishbach, Friedman, and Kurglanski 2003; Kruglanski et al. 2002). They found that temptations activate the goals with which they are seemingly in conflict. For example, one study involved participants with diet goals who were offered a prize of either an apple or a chocolate bar. Most of the participants that were primed with "diet" chose the healthier apple prize. Kruglanski et al. (2002) found that "for these dieting women, a tempting stimulus in the form of fattening delicacies was as effective as a direct dieting reminder not only in *activating* the dieting goal but actually in pursuing it behaviorally" (Kruglanski et al. 2002, 356). These interesting findings suggest that "temptations" do not always need to be actively and consciously overridden in order to pursue our higher priority goals. They can also automatically help remind

us of our goals through formed associations, where "temptations activate the 'larger' goals they are in conflict with, which, in their turn, tend to inhibit temptations" (Kruglanski et al. 2002, 356). Although this seems counterintuitive, it is an interesting finding that temptations could help us reinforce our long-term goals in certain cases, rather than acting as detractors.

Lastly, there is the choice model of self-control (Berkman et al. 2017; Berkman 2018), which argues, in contrast to dual-process models, that self-control is just another form of value-based choice (Inzlicht et al. 2021, 330). This model suggests that there is a single, unified, underlying valuation process where "various response options are assigned some subjective value and then a decision about which option to act upon is made though a dynamic integration of these competing values" (Inzlicht et al. 2021, 325). The value inputs are constantly changing and updating according to the context, whether there is uncertainty present, changing access to resources and attentional focus, etc. (Berkman et al. 2017, 423; Berkman 2018, 116). Therefore, this model presents a unique conception of self-control, one in which there are not two options competing, but rather, many options and alternatives in response to a fluctuating, highly dynamic context (Berkman et al. 2017, 423; Berkman 2018, 116).

Given the importance of self-control processes on a wide variety of behaviours, it is not surprising that there is so much research devoted to the topic. Models of self-control, as shown in this section, vary widely depending on their focus—whether this is a focus on conflict between long-term goals and short-term temptations and resource capacity (Baumeister et al. 1998), or competition between deliberative and impulsive systems (Hofmann, Friese, and Strack 2009; Baumeister and Heatherton 1996; Metcalfe and Mischel 1999). Alternatively, I also examined models less focused on conflict, instead focusing on how we monitor our goals and outcomes

(Carver and Scheier 1998), how our goals are represented (Kruglanski et al. 2002), or how our values are integrated to guide our goal-directed behaviour (Berkman et al. 2017).

Some of the newer, less traditional theories of self-control that reject dual-process accounts can provide us with models of self-control that can account for a more central role for emotion and motivation in self-control. This is more in line with some of the recent evidence showing that emotion and motivation do make important contributions to our control processes that go beyond acting as distractions from our long-term goals. More of this evidence for the role of emotion and motivation will be discussed in the following section, where I will consider the neural mechanisms of cognitive control processes.

3.1.3 Cognitive Control

As previously discussed, many recent accounts of self-control processes have been shifting away from seeing goal conflict as the central guiding force behind self-control processes. However, conflict *is* seen to be central to cognitive control processes. The consensus within the cognitive control literature suggests that cognitive control processes are, in fact, activated by conflict (Nigg 2017). This section will briefly discuss some prominent definitions and models of cognitive control, how this process impacts our goal-directed behaviours, and how emotion and motivation influence cognitive control.

Cognitive control can generally be understood as our ability to flexibly adapt behaviours and redirect attention in relation to our goals in the highly dynamic and everchanging environment in which we live. There is a great deal of research devoted to this process, and therefore many proposed definitions of cognitive control. Inzlicht et al. (2021) has defined cognitive control as "the flexible allocation of attention in the service of goal-directed behavior

in the face of more habitual or immediately compelling behaviors" (Inzlicht et al. 2021, 321). Another definition of cognitive control from Botvinick and Cohen (2014) defines it as the "real-time optimization of responses (emotion, attention, motor) to a goal" (Nigg 2017, 370; see Botvinick and Cohen 2014, 1260). And lastly, a definition from Pessoa (2009) defines cognitive control as a "set of functions…which are needed when non-routine behaviors are called for—namely, when 'control' is required. These functions are thought to confer behavioral flexibility and context dependency to complex behaviors" (Pessoa 2009, 160).

Pessoa (2009) goes on to explain that cognitive control includes, what he calls, "a host of adjustment processes, including perceptual selection, detection and resolution of conflict, and maintenance of contextual information" (Pessoa 2009, 160). This follows the prominent line of cognitive control research from Miyake et al. (2000), which states that cognitive control consists of three major processes, which are related but separable: mental set or task shifting ("shifting"), information updating and monitoring of working memory contents ("updating"), and inhibition of prepotent responses ("inhibition") (Miyake et al. 2000, 49).

While these three executive functions—shifting, updating, and inhibition—are thought to be separable[2], there is also plenty of evidence to suggest that the "mechanisms are shared across functions" (Pessoa 2009, 160). This means that the resources needed for these executive functions are not endless but rather are from a common shared pool. This idea of "'capacity sharing' is important for understanding information processing, as it leads to what is called 'executive competition'" (Pessoa 2009, 160). As stated by Pessoa (2009), "it is proposed that subcomponents of executive control are mutually interacting, such that resources devoted to one component will detract from those available to other components" (Pessoa 2009, 160).

[2] See (Miyake et al. 2000) for a detailed discussion of the unity versus diversity of executive functions debate.

As previously mentioned, cognitive control is generally understood to be activated by conflict. According to (Nigg 2017), this could be response conflict (where two responses have been primed, but only one can be enacted), perceptual conflict (where there is task-irrelevant perceptual information interference), cognitive conflict (where task-irrelevant associations and thoughts act as a hindrance to our choices or actions), or goal conflict (where an action will support one goal while interfering with another goal) (Nigg 2017, 370).

Several cognitive neuroscience models of cognitive control have focused on the importance of conflict and on monitoring systems that track response conflicts or prediction errors (Inzlicht, Bartholow, and Hirsh 2015, 128; see Botvinick et al. 2001; Holroyd and Coles 2002). One prominent model is the conflict monitoring theory (Botvinick et al. 2001), which suggests that there is a dedicated monitoring system that attends to conflict between co-activated, but mutually exclusive actions (Dignath et al. 2020, 194). This monitoring system is thought to work through a signal that is activated by the conflict and triggers changes in attention and control processes according to the "current task demands" (Dignath et al. 2020, 193). According to this view, the conflict monitoring system plays a crucial role in identifying conflicts and activating control processes in order to resolve these conflicts, thereby keeping goal-directed behaviour on track.

The anterior cingulate cortex (ACC) has been identified as a central brain region involved in this monitoring process (Botvinick et al. 2001). In particular, one area of the ACC—the dorsal anterior cingulate cortex (dACC)—has been shown to be a key node that integrates cognition and emotion (Rahman et al. 2021, 4765). Research has suggested that the dACC is consistently activated during conflict tasks, and seems to be crucial for monitoring conflicts, serving as an aversive signal that more control is needed to accomplish the task (Botvinick et al. 2001).

Initially, research on the functions of the ACC proposed a more segregated account, dividing the ACC into an "emotional" rostral section of the ACC and "cognitive" dorsal section of the ACC (See Inzlicht, Bartholow, and Hirsh 2015, 126 for a discussion of these approaches). However, much of the recent research has suggested that an integrationist account (Shackman et al. 2011) is more likely, where negative affect, pain, and cognitive control are served by overlapping regions of the dACC (Inzlicht et al. 2015, 126). This integration is thought to occur through the dACC's connectivity to both cortical and subcortical regions in the brain (Rahman et al. 2021, 4765). The overlapping hub of information has led to the dACC being implicated in a variety of different cognitive and emotional functions, such as decision-making, error processing, conflict monitoring, sympathetic arousal, as well as being central for negative emotional responses related to fear and anxiety (Rahman et al. 2021, 4765).

One key area of the dACC, the anterior midcingulate cortex (aMCC), has been a focus of some interesting recent research as well. Imaging studies have suggested an overlap between negative affect, pain, and cognitive control in the aMCC (Shackman et al. 2011). Additionally, anatomical studies show that the aMCC "constitutes a hub where information about reinforcers can be linked to motor centers responsible for expressing affect and executing goal-directed behavior" (Shackman et al. 2011, 154). This has led some researchers to hypothesize that activation in the aMCC could reflect a common, domain-general operation (Shackman et al. 2011). It is hypothesized that both cognitive control and pain (both activating the aMCC), act as an early warning system, allowing attention to be redirected and behaviour to be adjusted to avoid future errors (Shackman et al. 2011). Similarly, the researchers suggest that negative affect (e.g., fear, anxiety), like pain and cognitive control, is goal-directed and helps to flexibly coordinate responses and decrease the likelihood of future punishments (Shackman et al. 2011,

160). Additionally, Shackman et al. (2011) state that "demands for cognitive control are also thought to motivate new learning, such demands may serve as teaching signal that penalizes choices, strategies, or actions requiring greater control, promoting avoidance of cognitively taking actions in the future" (Shackman et al. 2011, 160).

Therefore, the hypothesis suggested to explain the domain-general functional contribution of the aMCC is that cognitive control, pain, and negative affect, engage in the same processes to solve the same problems brought about by contextual and environmental uncertainty. As further discussed by Shackman et al. (2011),

> Based on the data, we propose that aMCC implements adaptive control by integrating information about punishment (e.g. likelihood, magnitude) arriving from the amygdala, spinothalamic system, striatum, insula, and other regions in order to bias responding in situations where the optimal course of action is uncertain or entails competition between alternative courses. Outgoing control signals would presumably be sent directly to subcortical and cortical motor centers. Or control signals generated in aMCC and directed at the amygdala or lPAG might serve to resolve conflict between passive and active defensive behaviours. (Shackman et al 2011, 160)

In other words, the ACC is particularly important for integrating information from various sources, including motivational and emotional information through the ACC's various connections to the amygdala, orbitofrontal cortex, and insula (Pessoa 2009, 163). Additionally, as previously discussed, the ACC has been implicated in processes related to conflict detection, error processing and monitoring, and determining the benefits and costs of acting (Pessoa 2009, 164). This shows a wide-ranging impact of the ACC on control processes—many of which crucially involve emotional and motivational information. The idea that emotion and cognition are integrated in cognitive control processes has had a lot of support from various research groups (Shackman et al. 2011; Dreisbach 2006; Dreisbach and Fischer 2015; Chiew and Braver 2011; Pessoa 2009; Okon-Singer et al. 2015; Botvinick 2007; Inzlicht et al. 2015). There have

been many studies that have shown that emotion and motivation have a significant effect on cognitive control (Braver et al. 2014)[3].

Therefore, the connections between emotion, motivation, and control play an important role in when it comes to our decisions and actions. According to several influential accounts of control, the decision about when more executive control resources are needed is based on a type of cost-benefit analysis (Kool and Botvinick 2014), where the benefits are weighed against the cost of using up valuable control resources (Braver et al. 2014, 457). According to Braver et al. (2014), the ACC might act as a critical hub, connecting motivation with executive function, playing a central role in the computation of the expected value of control (Braver et al. 2014, 457).

According to this view of the role of the ACC in conflict monitoring mechanisms, conflict is registered as an aversive signal, which would bias behaviour toward actions that minimize conflict, or the possibility of future conflict (Botvinick 2007, 359). As discussed by Botvinick (2007),

> If this hypothesis is correct, it would establish a clear alignment between the conflict-monitoring and the outcome-evaluation/decision-making perspectives on ACC function. Conflict would be placed within the larger set of value-laden outcomes registered by the ACC. In particular, like monetary loss, negative feedback, social exclusion, and pain—other occurrences registered within the ACC—conflict would weigh as an aversive or costly event. Furthermore, this weighting would have a direct impact on decision making, influencing subsequent behaviour though domain-general avoidance-learning mechanisms which are already known to involve the ACC. (Botvinick 2007, 359)

The idea that conflict (as well as punishments and errors) is experienced by us as aversive and unpleasant, ultimately leading to avoidant behaviour, is supported by an increasing amount

[3] See (Braver et al. 2014) for a more detailed discussion of the impact of motivation and emotion on cognitive control processes.

of research (Okon-Singer et al. 2015; Botvinick 2007; Kool et al. 2010 ; Dreisbach and Fischer 2015; Shenhav and Buckner 2014; Dignath et al. 2020; Inzlicht, Bartholow, and Hirsh 2015). Based on these insights, various researchers have proposed an "affective-control framework" (Dreisbach and Fischer 2015; Inzlicht et al. 2015; Saunders et al. 2017; Van Steenbergen 2015; Dignath et al. 2020) that suggests a central role for conflict-triggered affective control (Dignath et al. 2020, 196).

In addition to the ACC as a central mechanism for conflict signaling, another key hub supporting control processes is the dorsomedial prefrontal cortex (Braver et al. 2014, 456). As discussed by Braver et al. (2014), a "general consensus view is that the ACC and adjacent dorsomedial PFC serve an evaluative role in monitoring and adjusting levels of control…potentially in response to motivational variables" (Braver et al., 2014, 456-457).

The connection between motivation, emotion, and cognitive control, and the "affective-control" framework more generally, have opened the door for recent theories suggesting cognitive control should be understood as an emotional process (Inzlicht et al. 2015; Dignath et al. 2020). The 'affective-signaling hypothesis' (Dignath et al. 2020), for instance, can be understood as building upon conflict monitoring theories. This account is based on the premises that (a) conflict and errors elicit an affective reaction, (b) performance monitoring not only detects conflicts and errors, but also importantly detects the affective reactions to these errors, and lastly, (c) that this affective response to conflict and errors are the main drivers of the changes in adaptive control (Dignath et al. 2020, 196). This last point is further argued by Inzlicht et al. (2015), arguing that negative affect is a central, initiating aspect of cognitive control (i.e., that affect plays a *causal* role in control and is not just a side-effect or reaction to the process) (Inzlicht et al. 2015, 126).

The role of affect then, is integral to understanding how goals, motivation, and cognitive control are linked in guiding behaviour choices. Motivation—the "internal state that drives behavior toward a rewarding goal or end point and away from undesirable or punishing outcomes" (Inzlicht et al. 2015, 128)—is a crucial process that helps support our goal-directed behaviour choices. According to this view, initially the conflict between goals triggers cognitive control mechanisms by means of an aversive affective signal. However, motivational information is also important in this process because some goals will be inherently more valuable than others (Inzlicht et al. 2015, 128). The goals that are more valuable are more likely to engage these control processes (Inzlicht et al. 2015, 128). In other words, the more highly a goal is valued, the more highly motivated a person would be to try and accomplish the goal (Inzlicht et al. 2015, 128). Therefore, according to this view, motivation plays a central role in cognitive control (Inzlicht, Bartholow, and Hirsh 2015, 128; see also Chiew and Braver, 2011; Braver et al., 2014). Inzlicht et al. (2015) further discusses how motivation can be understood to direct cognitive control below:

> The motivation and willingness to engage control efforts depends on a trade-off between the expected payoff from a controlled process (e.g. acquiring some desirable reward or avoiding some undesirable punishment) and the perceived cost of engaging control in terms of cognitive effort (Shenhav et al. 2013). When these factors lead individuals to feel motivated, cognitive conflict is particularly salient, resulting in considerable task-related negative affect when control fails or is at risk of failing (Lindström et al. 2013). This negative affect is not merely output, however, and can result in goal reprioritization and the mobilization of further goal-directed actions (Carver and Scheier 1998). Emotion, in our view, is consequently both an output of motivation and an input to the execution of goal-directed behaviour…Motivation, in short, directs cognitive control (Pessoa 2009; Chiew and Braver 2011), with emotion as output of and input to processes that engender control. (Inzlicht et al. 2015, 128)

In sum, the affective accounts of cognitive control suggest that instead of the traditional view of affect competing with cognitive control, affect may instead be an essential, and

potentially even a causal component of cognitive control (Inzlicht et al. 2015, 130). Viewing cognitive control as an emotional process raises interesting new questions, especially in light of the previously discussed role that emotions play in self-regulation and self-control processes. If emotion and motivation play a central role in many control processes, such as self-regulation, self-control, and cognitive control, then we should move away from frameworks that pit emotion and reason as competing for dominance over our behavioural choices.

Furthermore, bridging the research on control with the research on decision-making has provided some interesting insights in areas where the two processes connect (Frömer and Shenhav 2022). For instance, we know that goals play a central role in decision-making; however, cognitive control research can provide further information on how goals are regulated and updated, as well as how goals guide selection processes in decision-making (Frömer and Shenhav 2022). Cognitive control research can also provide insights to other aspects of decision-making, such as how decision-making processes can be affected by other factors such as personal biases or uncertainties in the environment (i.e., is there enough evidence to make a good decision, whether the choice options are good or bad), as well as how we monitor the choices we have made in the past, present, and future, and whether our goals need to be adjusted based on these choices (Frömer and Shenhav 2022).

Frömer and Shenhav's (2022) article provides interesting suggestions about where and how the decision-making processes and cognitive control processes interact in the brain. First, there has been a quite a bit of cognitive control research suggesting that goals affect the way we process stimuli (Frömer and Shenhav 2022, 3). When considering the effect of cognitive control processes on value-based decision-making, Frömer and Shenhav have argued that cognitive control acts on value representations analogously to how it acts on perceptual information in that

it directs attention to the attributes of our choice options that are weighted with the most value (Frömer and Shenhav 2022, 3).

They further explain that the evidence suggests that the dorsolateral prefrontal cortex (dlPFC) region of the brain maintains the goal representation and signals stimulus processing regions to prioritize certain attributes over others when trying to decide how to respond to choice options (Frömer and Shenhav 2022, 3). Therefore, the dlPFC, in this role, modulates value-related activity in the ventromedial prefrontal cortex (vmPFC) in the same way that it plays a modulating role in directing attention toward goal-relevant information in cognitive control research (as shown in cognitive control tasks such as the Stroop task) (Frömer and Shenhav 2022, 3). Frömer and Shenhav (2022) claim that "under this view, it is natural to think of control as directing the flow of reward information with the value network, such that neural signals within vmPFC come to signal how rewarding an option is based on one's current goals" (Frömer and Shenhav 2022, 3).

As reviewed in Chapter Two, evidence accumulation models have been helpful in explaining the processes underlying value-based decision-making. Evidence accumulation models show how noisy evidence is neurally accumulated over time, what is called a "decision boundary," where a prespecified level of evidence is reached. Once this threshold is surpassed, the decision is made (Frömer and Shenhav 2022, 5).

Frömer and Shenhav (2022) argue that there are quite a few aspects of this decision-making process where control mechanisms are exerted. For example, they suggest that many of the evidence accumulation models assume that both the input and decision-bound remain constant throughout the decision-making process, as well as the amount of evidence needed to reach the decision (Frömer and Shenhav 2022, 6). However, Frömer and Shenhav (2022) point

out that this, in fact, may not be the case. Instead, they present evidence to show that "control might be recruited to (1) dynamically adjust the inputs to evidence accumulation, (2) the threshold for decision (decision-bound), and (3) and/or to inhibit biases towards particular options" (Frömer and Shenhav 2022, 6). In other words, we exert control over the flow of information needed as evidence input for our decisions, we exert control over how much evidence we need for reaching our decision thresholds, and lastly, we also exert control over which option we end up choosing due to biases we may already have developed (Frömer and Shenhav 2022, 6). The recruitment of control in the various aspects of decision-making has been shown through the consistent activation of the dACC—the same area recruited for classic control tasks (Frömer and Shenhav 2022, 7). Incorporating research on the intersection between decision-making processes and control processes can provide insights into how these processes work together to represent our goals and guide goal-directed behaviour.

Overall, the last few decades have provided a wealth of research on cognitive control processes, as well as how cognitive control is affected by emotion and motivation processes. A key insight is the finding that conflict triggers an affective signal, which in turn activates cognitive control processes—a process that seems to rely on the involvement of several key areas of the brain like the dACC, aMCC, and dlPFC. A causal role for emotion in cognitive control theories have changed the framework quite dramatically, with interesting implications for self-control dilemmas that I will explore in the next section.

By examining the processes for self-regulation, self-control, and cognitive control, my aim was to clarify the ways in which they each contribute to our goal-directed behaviour. Additionally, despite occasional overlap, the empirical research suggests that they should in fact be considered distinct processes, with their own unique contributions. Importantly, the research

reviewed also highlighted all the various ways that emotion and motivation impact these control processes, therefore supporting my argument that emotional and motivational information play a central role in our control processes. This is an important point that at times gets lost in discussions of self-control and is understated in philosophical discussions of self-control. The next section builds on this point and considers the way self-control dilemmas have been traditionally discussed in the self-control literature.

3.2 Control as a Duality: Emotion in Self-Control Dilemmas

As previously discussed, self-control dilemmas within the empirical literature have traditionally been framed as a conflict between a long-term goal versus a short-term temptation. Interestingly, this is also paralleled in the philosophical autonomy literature, where instances of people struggling with overcoming tempting desires are common examples of the struggle to govern oneself autonomously. As previously discussed in the self-control section above, presenting the dilemmas this way posits self-control as an active, conscious process of overcoming temptations—the temptation must be overcome to achieve one's goals and to be successful at self-control.

This type of dilemma is also prevalent in the dual-process models of self-control. Berkman (2018) refers to the dual-process models as "oppositional models," where slow, reflective mental processes support our long-term goals and compete with the fast, impulsive processes representing the more tempting course of action (Kahneman 2011; Heatherton and Wagner 2011; Hofmann et al. 2009) (Berkman 2018). On a dual-process account, the self-control process occurs when, in response to this competition, we try to consciously and effortfully overcome the tempting impulse and stay on track with our long-term goals (Berkman 2018, 114).

Philosophy has a similar tradition of using this type of dilemma between our rational goals and our more hedonic temptations as representative of self-control. As Berkman and colleagues (2017) also point out, this idea of self-control is not new and its roots can be traced all the way back to Plato, if not even further. For instance, in Plato's *Phaedrus*, we find Plato's discussion of the soul using the famous metaphor of the charioteer and his two winged horses (Plato 1997, 506–56). The horse on the right is good, noble, and obedient. The horse on the left is unruly and untamed, and represents the uncontrolled drive toward hedonic pleasures (Belfiore 2006). The charioteer represents the guiding principle in the soul (Belfiore 2006, 190), which controls the impulses of both horses, pulling the charioteer in two different directions. For Plato, the three parts of the chariot (charioteer and his two horses) must work together in harmony (*Phaedrus,* 253e).

We can see similar normative assumptions in both Plato's chariot metaphor and the opposition models of self-control. That is, the idea that the rational, "goal-promoting" choice is in competition with the impulsive, irrational temptations, which are "goal-preventing" (Berkman et al. 2017). The central idea in self-control dilemmas is the same central idea as Plato's charioteer metaphor—the "bad" horse, with its hedonic, irrational impulses, must be controlled. In self-control dilemmas, it is presented as a competition between the two opposing options (e.g., our reflective long-term goal, versus our irrational short-term reward), but importantly, it is also presented as a duality with only two options available. However, as various self-control researchers have recently pointed out (Inzlicht et al. 2021; Milyavskaya et al. 2019; Berkman 2018; Berkman et al. 2017), this assumption needs to be questioned. It is not always the case that we only have the two choices, a competition between the "good" goal or the "bad" temptation (i.e., the classic struggle between the devil and angel on our shoulders). Even if self-control feels

like a duality between the "good" goal and "bad" temptation, there can be many alternative courses of actions related to self-control (Milyavskaya, Berkman, De Ridder, 2019).

The example used by Berkman (Berkman et al. 2017, 422) is of a dieter trying to decide between ordering a salad or a hamburger for lunch. One option promotes the dieter's long-term goal of weight loss, the other option relates to the dieter's hedonic temptation/urge to order the tastier option. In this case, the traditional self-control oppositional models assume that the temptation must be effortfully inhibited in order to prevent the temptation from overriding the long-term goal of the dieter (Berkman et al. 2017, 422). However, it has recently been argued that there are, in fact, many other alternatives to deal with this type of self-control dilemma[4] (Berkman et al. 2017). For example, the dieter can choose to focus their attention on a particularly tasty aspect of the salad (e.g., the salad dressing), or focus on the satisfaction they will feel by accomplishing their weight loss goal (Berkman et al. 2017, 422). There may also be other options beyond the salad or burger, one that might be healthy and extremely tasty.

There are also proactive strategies to eliminate these kinds of temptations before the dilemma even occurs, such as choosing a restaurant with only healthy lunch options (Berkman et al. 2017, 422)[5]. Therefore, there are many other viable self-control strategies that go beyond effortful inhibition, and more choices available than just the two "good" versus "bad" choices. This version of self-control does not translate into the self-control conflict between one long-term goal versus one distracting temptation, a temptation that must be consciously and effortfully inhibited. Instead, an expanded self-control dilemma would have many different options with which to choose and different pathways to resolving a goal conflict (Berkman et al. 2017). This

[4] Fujita (2011) first argued that we need to expand the way we look at self-control.

[5] The alternative strategies are more thoroughly discussed in (Duckworth et al. 2016).

has implications for the way understand self-governing our choices, moving us away from the paradigm of reason being in control of our "lower" emotions and impulses.

Despite the initial popularity of dual-process models, with early studies supporting increased activity in the PFC when completing self-control tasks, newer research suggests that dual-process models may not be supported by much evidence. Recent studies looking at the brain activity in a self-control situation have shown both increased activity in the lateral PFC and decreased activity in the lateral PFC (Kelley, Wagner, and Heatherton 2015, 398). This mixed result does not support the dual-process position, which should show a consistent increase in top-down lateral PFC activity coupled with a decrease of activity in bottom-up reward networks. Due to these mixed results, it has been suggested that the studies in self-control that have shown decreases in lateral PFC activity does not necessarily indicate a person that lacks self-control, like initially thought (Kelley, Wagner, and Heatherton 2015, 398). Decreases in PFC activity can equally be a measure of someone lacking self-control and someone with exceptional self-control. That is, a person with excellent self-control strategies will likely avoid the temptations to begin with, without ever needing to effortfully inhibit any temptations. If this is the case, it would explain the lack of activation in the lateral PFC in a person that is otherwise quite successful at managing self-control tasks. The lateral PFC activation is only present when a particular type of highly effortful inhibition of temptations is required, hence the mixed results of the studies (Kelley, Wagner, and Heatherton 2015, 398).

Another line of evidence that points us away from dual-process models comes from the research discussed in the previous chapter, on value-based decision-making processes (Kable and Glimcher 2007). The research from Kable and Glimcher (2007) suggests that activity in regions of the brain that were thought to be involved in only automatic reward responses, is actually

involved in computing the value of both controlled and impulsive choices (Berkman et al. 2017, 424). There is also evidence showing that activity from different brain areas tracks the value of distinct attributes related to our choices (i.e. gains and losses, emotional and utilitarian benefits of moral actions, option's value for self and others, and the value of waiting for a better outcome (Berkman et al. 2017, 424).

Importantly, it was demonstrated by (Clithero and Rangel 2014) that these specific representations of the various attributes converge in the traditional value-related areas of the brain, such as the ventral striatum, ventromedial prefrontal, and orbitofrontal cortices—areas in which the overall subjective value of choice options are tracked (Berkman et al. 2017, 424). According to Berkman et al. (2017), this research suggests more of a value integration process for self-control, rather than a competitive, inhibitory relationship (Berkman et al. 2017, 424). As further discussed by Berkman et al. (2017),

> This architecture suggests that self-control operates as a valuation process rather than a battle between different systems. Dual-systems models generally postulate that systems representing long-term attributes and hedonic considerations compete to inhibit each other, with the winner driving behaviour. Yet neural evidence for this kind of reciprocal inhibition is scarce (Hutcherson, Bushong, and Rangel 2015; Kelley, Wagner, and Heatherton 2015). In contrast, value signals in regions like the vmPFC track choices regardless of whether that choice is patient or impatient, healthy or unhealthy, charitable or selfish. Self-control outcomes are determined by the relative degree to which the value of all attributes are reflected in vmPFC (Hutcherson et al. 2015; Kable and Glimcher 2007). (Berkman et al. 2017, 424)

Focusing on self-control as a more integrated, valuation process can provide us with a broader perspective of self-control that incorporates inputs from various regions of the brain. As Berkman et al. (2017) further discusses,

> Control networks such as lateral PFC contribute to self-control by influencing the weights given to different attributes in the value integration process, rather

> than by inhibiting other regions (Hare et al. 2011). Thus, self control outcomes emerge organically from the operation of a single, integrative system with input from multiple regions rather than antagonistic competition between two processes. (Berkman et al. 2017, 424)

The most interesting conclusion from Berkman et al. (2017) is that if self-control dilemmas are just value-based choices, like all other choices between alternative options, then there is nothing special about these kinds of dilemmas that have dominated both the empirical self-control literature as well as the philosophical autonomy literature. According to this view, self-control dilemmas where you feel like you're being pulled in two different directions, or dilemmas where you feel like one choice over another is very motivationally difficult and effortful, do not feel that way because they are a special sort of normative dilemma, nor does it involve our "will" attempting to gain control over temptations or impulses. It might just be another type of value-based choice that involves a particularly difficult conflict to resolve. If our self-control dilemmas are just ordinary value-based choices, a situation that initially seems like a choice between a long-term goal and a short-term temptation, is going to follow the same valuation process as any other value-based decision that you need to make. Meaning, all the options you have available undergo evaluation and are weighted in an attempt to calculate which is the best option. The weighting is based on the various attributes present in your available options, and do not necessarily involve the conscious, reflective, inhibition of temptations.

This view of self-control as a value-integration process, where short-term and long-term affective consequences (costs) of the conflicting options are weighted in a common subjective value currency in the vmPFC is supported by a recent study that attempted to test "real-life" self-control choices (Krönke et al. 2020). This study set out to test the ecological validity of the traditional, but highly artificial, self-control laboratory tasks—addressing concerns that the laboratory self-control dilemma tasks artificially narrow choices which may not be reflective of

how we make value-based choices in real life. There have also been concerns over these sorts of laboratory studies having poor associations between the tasks and self-reported self-control in participants, as well as poor predictive power and a general lack of reliability (Krönke et al. 2020, 268).

As opposed to the dual-process theories, with a "reflective" system encoding long-term goals while suppressing the "impulsive" system which encodes short-term value (Krönke et al. 2020, 269), Krönke et al. (2020) hypothesized that self-control in real life settings might involve deciding between conflicting choice options based on a dynamic value integration process where the subjective value of each of the alternative options is calculated through the integration of different short-and long-term gains and costs related to the options (Krönke et al. 2020, 269). According to Krönke et al. (2020), "the resulting integrated value signal reflects the individual's preferences at the moment of decision and will determine the enactment of the most valued option" (Krönke et al. 2020, 269). Thus, this approach understands self-control differences in individuals as differences in the weight with which long-term consequences are calculated in this value-integration process, rather than self-control as a conscious, reflective process of inhibiting impulsive, short-term rewards (Krönke et al. 2020, 269).

The results of this study showed that activity in the vmPFC plays a central role in the modulation of a common value signal by the anticipation of long-term consequences of the choice options (Krönke et al. 2020, 276). The vmPFC was also shown to reflect the subjective value related to a wide variety of different goals and desires, in both laboratory self-control choices, as well as real-life self-control choices (Krönke et al 2020, 276). When discussing the results of their study, Krönke et al. (2020) state,

> [Their] results provide direct evidence that not only relatively inconsequential choices in laboratory tasks but also self-control in real-life conflicts reflect

> individual differences in how anticipated long-term consequences are weighted into integrated neural representations of subjective value. (Krönke et al 2020, 276)

In terms of the role that cognitive control networks play in self-control, Krönke et al. (2020) suggests that this activity may be related to the "top-down adjustment of the relative weight with which long-term outcomes and their anticipated affective consequences are integrated into a common value representation in the vmPFC" (Krönke et al. 2020, 277). Importantly, what this study provides is evidence of a direct correlation between affective value and self-control processes, in both laboratory and real-life self-control choices, therefore adding support to self-control models that consider self-control processes as more of an emotional process than traditionally understood. This study also supports Berkman et al. (2017) model of self-control as value-based choice—where the costs and benefits (long- and short-term) of choice options are weighted in a value-integration process, rather than two systems operating in competition with one another.

Going back to the earlier example of the dieter, self-control experiments are often set up to involve a choice between only two options, one that is tied to a "cold," long-term goal and another that is a "hot," short-term temptation. This is only one type of decision that people face in day-to-day life. In real life, people are complex, often have many competing long-term goals tied to many complex values. When we artificially narrow self-control dilemmas down to these types of two-choice, hot versus cold self-control dilemmas, it prevents us from fully understanding the role that self-control processes play in our life choices, and in particular, in our autonomous choices. It also moralizes the conflicts between long-term goals and short-term temptations by suggesting the deliberative option is morally better as it helps us achieve our more reflective goals rather than being distracted by impulsive temptations leading us astray. The

reality of it is that these choice alternatives will fluctuate, and some will support some actions that may be good or bad for a person. Furthermore, at times the short-term temptation may do a better job supporting long-term goals then the more deliberative choice option. The assumption that the long-term, cold choice supports goals, whereas the short-term temptation prevents goals, may only be true when the dilemmas are artificially constructed to be that way by experimenters (Milyavskaya, Berkman, and De Ridder, 2019).

Conclusion

The aim of this chapter was to examine some of the prominent neuroscientific research of control processes to better understand self-control and its related concepts. One main goal of this chapter was the clarification of the differences between the terms self-regulation, self-control, and cognitive control. Despite occasional overlap, the empirical research suggests that they should be understood as distinct processes. Importantly, the research discussed in this chapter demonstrates the integration of emotional and motivational information in the three types of control.

Self-regulation was defined as a broad umbrella concept, which includes various processes related to goal-directed behaviour, such as how we choose which goals to pursue, how we monitor, plan, and implement those goals, and how we shield goals from competing actions and thoughts. Self-regulation is understood to involve various neural networks, such as the cingulo-opercular network (including the inferior regions in the lateral PFC, regions of the ACC, insula, and other nodes), the fronto-parietal network (including superior regions in lateral PFC like the dorsolateral PFC, the posterior parietal and inferior temporal cortex, and subcortical connections), as well as cross-talk with other networks such as the default mode network

(including the ventromedial PFC), the noradrenergic, locus-coeruleus, 'alertness network', and subcortical regions (in particular, the thalamus and hippocampus) (Nigg 2017, 366). The fact that many networks and areas of the brain support the self-regulation process demonstrates that it is a broad concept that encompasses many different aspects of regulating our thoughts, emotions, actions, and behaviours.

Self-control, on the other hand, is a narrower concept—at times included under the umbrella concept of self-regulation. Self-control was initially defined by conflict, as an inhibitory process that is initiated when a person has conflicting response options between a long-term goal and short-term temptations. However, newer research (Fujita 2011; Galla and Duckworth 2015; Gillebaart and de Ridder 2015; Hennecke et al. 2019; de Ridder et al. 2011; Gillebaart 2018; Duckworth et al. 2016; Berkman et al. 2017; Berkman 2018; Inzlicht et al. 2021) has expanded the working definition of self-control as a process that includes all methods of resolving conflicts between competing goals (i.e., methods that do not require conscious effort, inhibiting something undesired or enacting a desired outcome, methods of reactive inhibition or proactive conflict resolution) and is simply considered to be the process of pursuing one goal over another when the two come into conflict (Inzlicht et al. 2021, 321). The goals in conflict do not need to be a long-term goal and a short-term goal—they can also be long-term goals versus long-term goals, or short-term goals versus short-term goals. Some recent models have suggested a key role for emotion in self-control processes, moving away from traditional accounts that view emotion as a hindrance to self-control. By widening the definitions of self-control, we can more fully understand how processes integrate to support this function which has been shown to be crucial for overall well-being.

And lastly, research on cognitive control, has also recently focused on the emerging connections between emotion, motivation, and cognitive control. Several theories have noted a central role of affect in cognitive control processes—with an aversive signal in the dACC acting as a catalyst for changes in adaptive control.

The role of emotion and motivation in our various control processes was a key insight that emerged from the recent empirical literature. Much of the research focuses on the various ways in which emotion and motivation impact control. However, some of the research suggests a causal role for affect in self-control processes, as well as self-regulation and cognitive control processes. An appreciation of the emotional/motivational foundation of control-processes has important implications for the way we view control and the concept of autonomy. Additional research on the role of emotion in self-control will help us further understand the role of emotion in autonomy.

The role of emotion and motivation in self-control has also contributed to an expansion of our conception of self-control dilemmas. Self-control has long been overly relegated to a dilemma that pits deliberative, long-term goals against emotional, short-term rewards. This type of dilemma of course, can be a type of conflict that we face. However, it is not the only type of self-control dilemma that we deal with on a day-to-day basis, and therefore is not in and of itself, reflective of how we use self-control in our daily choices and behaviours. This shift away from these types of paradigmatic self-control dilemmas has direct implications for personal autonomy, as these examples are prominent in the autonomy literature. These examples are commonly used as an example of someone struggling to govern their desires. I will explore this idea further in later chapters, and in particular, when looking at Frankfurt's use of self-control dilemmas. Overall, in this chapter I argued that we should consider moving away from this overly narrow

conception of self-control and explored some recent evidence for alternative accounts of self-control that support a broader conception of self-control as an affective, integrated-value process.

Chapter 4

The Neuroscience of Voluntary Actions and the Experience of Agency

Introduction

Related to the control literature discussed in the previous chapter, an interesting subsection of research has emerged that focuses on intentional, voluntary, self-generated actions (Libet et al. 1983; Libet 1985; Haggard and Eimer 1999; Brass and Haggard 2007; 2008; Haggard 2008; 2017). This research initially stemmed from the well-known Libet experiments on conscious free will in the 1980s, where Libet claimed his experiments demonstrated that our actions precede our conscious awareness of a voluntary decision to act. While the neuroscience of autonomy is still rather underexplored, there has been decades of research looking into the neuroscience of free will since the experiments by Libet and colleagues. Although this is slightly outside of the scope of this dissertation, some of the research on the neuroscience of free will touches on important ideas related to personal autonomy. Therefore, it is worth looking at some of the studies that might bear on the topic of autonomy.

I first examine some of the research on the neuroscience of free will, primarily the Libet experiments. Following this, I look at two branches of research that have stemmed from these early experiments on free will: the neural correlates of voluntary action and intentional inhibition. This research is a small segment of a much larger literature on intentional action. I aim to review some of the research that most closely correlates with ideas that are discussed in relation to autonomy. The research on voluntary action and intentional inhibition looks at the neuroscience of voluntary motor action decisions, considering what causes a person's internal, voluntary action choices and how one chooses to inhibit these actions. The idea is that this type

of research captures something important about the way we think about "volition," and our capacity to causally determine our own actions.

Relatedly, there is also ample research on the neural mechanisms underlying the conscious experiences of agency and ownership over our voluntary actions. I will examine some of this literature in the second section of this chapter. This research looks at how our brain creates the conscious experience of feeling like our actions have been internally caused, as well as feeling a type of control and ownership over our actions. Throughout this chapter, I also discuss a few different types of conditions which produce so-called "autonomy disturbances." These conditions have a profound impact on the ability to voluntarily control one's actions and effect how one's actions are experienced.

Overall, this research, while focused more on free will, can also provide some insights on ideas that are central to personal autonomy. As I will discuss further, the empirical research discussed in this chapter indicates that many of our voluntary action choices are based on the emotional and motivational value information tied to our options, much of which is applied automatically and nonconsciously. I argue that this is quite different than traditional ideas of conscious, deliberative decision-making in the philosophical autonomy literature. In later chapters, I will reflect on how some of these experiments on voluntary action choices and the experience of agency may impact related ideas raised in Korsgaard's and Frankfurt's approaches to autonomy.

4.1 Libet's Experiments and the Neuroscience of Free Will

A series of experiments in the 1980s by Libet and colleagues established an area of research focused the neuroscience of free will. The central finding in these experiments seemed

to show that our brain was already unconsciously preparing for a specific action approximately 350ms before we even had a conscious urge to act. This led Libet to question our common understanding of free will as a conscious intentional decision (Libet et al. 1983). Libet's experiments were designed to test subjects performing endogenous (internally caused), voluntary movements and the conscious awareness of their intention or urge to act. Subjects in one group were asked to perform a quick movement of their right fingers and/or wrist at a time of their own choosing. Another test group was similarly asked to voluntarily move their right fingers or wrist, but this group was given further instructions "to let the urge to act appear on its own at any time without any preplanning or concentration on when to act" (Libet et al. 1983, 625). Both groups had to also self-report based on a revolving clock, the time that they were consciously aware of their intention to move their fingers/wrist (Libet et al. 1983).

Libet and colleagues then examined the timing of the subject's awareness of their intention to move and compared this with the neural signals that reflected the initiation of action (EEG scalp-recorded cerebral activity measured as the readiness-potential or RP, representing motor preparation in the brain originating from the supplementary motor area (SMA)) (Libet et al. 1983). Libet claimed that their data showed that the time of the onset of readiness potential occurs (on average) 350ms or more before the subjects were aware of an intention to move. Based on these initial trials, Libet concluded that the results provided evidence that actions are not consciously initiated and that this could potentially be interpreted as a lack of conscious free will as it is usually known (Libet et al. 1983).

Later experiments by Libet (1985) raised another interesting point about our current understanding of conscious free will. In Libet's (1985) experiment, he replicated the earlier findings of RP onset preceding the subject's conscious awareness of an intention to act.

However, he also found that there may still be room for the subject's conscious control over the actual performance of the final movement. According to Libet (1985),

> I propose the thesis that conscious volitional control may operate not to initiate the volitional process but to select and control it, either by permitting or triggering the final motor outcome of the unconsciously initiated process or by vetoing the progression to actual motor activation. (Libet 1985, 529)

He proposes here that there is a small window of time, between the subject becoming consciously aware of the intended action choice, where the subject has an ability to consciously trigger or veto the action that has already been prepared by the unconscious brain activity (Libet 1985, 538). Therefore, while a subject may not be consciously *initiating* the generation of their action, they may still be able to assert their "conscious free will" by *allowing* the action to either proceed or be cancelled.

Following Libet's line of research has led to a vast amount of neuroscientific studies on, what has now been termed, the "neuroscience of volition" (Haggard and Eimer 1999; Brass and Haggard 2007; 2008; Filevich et al. 2012; Lynn et al. 2014; Fried et al. 2017). I will take a closer look at two areas specifically, voluntary actions and intentional inhibition. This research examines the causal source of our actions—whether these actions are internally self-generated. Additionally, this research examines Libet's hypothesis of whether or not we can consciously inhibit our actions. Although these lines of research are not exactly focused on personal autonomy, they still have some interesting parallels and contribute to an understanding of the way the brain produces behaviours related to autonomy.

4.1.1 Brain Circuits for Voluntary Action

Haggard (2008) provides a detailed summary of the main cortical motor circuits that he has found to be involved in voluntary action. The circuits converge on the primary motor cortex (M1), which transmits motor commands to the spinal cord and muscles (Haggard 2008, 936). The motor cortex receives inputs from the two main circuits that underlie voluntary and stimulus-driven actions: the basal ganglia-pre-SMA circuit and the parietal-premotor circuit (Haggard, 2008).

The first key input to the M1 comes from the pre-supplementary motor area (pre-SMA), which receives its inputs from the basal ganglia (BG) and prefrontal cortex. The pre-SMA forms part of the wider cognitive-motor network that includes the premotor, cingulate, and frontopolar cortices (Haggard 2008, 936). As previously discussed throughout Chapter Two and Three, the BG (through its input to the pre-SMA) seems to again play a significant role in the initiation of action (Haggard, 2008). A 2004 study (Loukas and Brown 2004), showed that signals can be recorded in the BG up to two seconds before a forthcoming voluntary response, meaning BG activity precedes recorded readiness potentials, or movement preparatory action in the supplementary motor areas (Haggard 2008, 936). As Haggard mentions, this makes it tricky to pinpoint exactly where the first initiation of voluntary action begins. “The brain’s circuits for voluntary action might consist of loops rather than linear chains that run back to an unspecified and uncaused cause (‘the will’)” (Haggard 2008, 936). If we do not know how far back to go, or if a linear approach to understanding the initiation of voluntary action is not appropriate, then asking when and where endogenous voluntary action is initiated may in fact, be the wrong question.

What we do know is that the basal ganglia, through its subcortical loops, integrates a wide range of signals, modulated by dopaminergic input, to coordinate appropriate actions for a particular context that takes into consideration reward value and the motivation to act. According to Haggard, from this viewpoint, "voluntary action is better characterized as a flexible and intelligent interaction with the animal's current and historical context than an uncaused initiation of action" (Haggard 2008, 936).

The second key cortical network is the parietal-premotor circuit, which converges on the primary motor cortex (M1). According to Haggard, this circuit plays a role in the immediate sensory guidance of actions (Haggard 2008, 936). Information from sensory cortices is sent to the parietal lobe, which in turn, projects to the premotor cortex, which finally transmits these inputs to the M1 (Haggard 2008, 936). The parietal-premotor circuit underlies actions that are object-oriented, such as grasping objects in your current environment. However, Haggard also claims that this circuit may underlie some other types of voluntary actions. For example, single-unit recordings from the primate intraparietal sulcus in tasks in which the animal must choose between equally rewarding alternatives show that neurons in the intraparietal area (LIP) might encode the specific behavioural outcome that the animal has chosen (Haggard 2008, 936). This research on the neural activity in the intraparietal area (area LIP) was previously discussed in Chapter Two, where we saw that activity in this area was central to both perceptual decision-making and value-based decision-making. Therefore, as Haggard claims, the two circuits (the BG-pre-SMA circuit and the parietal-premotor circuit) may help implement different kinds of decisions of relevance to voluntary actions (Haggard 2008, 936).

4.1.2 Brain Circuits for Intentional Inhibition

The term intentional inhibition "refers to the voluntary and internal decision to withhold from executing a prepotent action tendency"—that is, the ability to suppress an ongoing, pre-planned movement based on an internally generated voluntary decision to stop (Lynn et al. 2014, 247). Brass, Haggard, and colleagues, attempt to distinguish this particular type of self-control (intentional inhibition) from the inhibitory self-control research previously discussed in Chapter Three (See Aron 2011 for a review of the inhibitory control literature) on the grounds that intentional inhibition is endogenous (or internally caused) and more characteristic of what we would consider "volition."

Haggard and Brass claim that the inhibitory control network discussed in the last section, consisting of the network involving the IFG, pre-SMA, basal ganglia, and STN, is a type of inhibitory self-control that is guided by an external signals or external events (Lynn, Muhle-Karbe, and Brass 2014, 252). Instead, Brass and Haggard (2007) focus on specifically investigating what happens when participants voluntarily (free from external signals or external events) choose to start or stop an action. For Brass and Haggard, the voluntary decision of whether or not to act is a central aspect of intentional action and particularly important because it allows agents to form intentions without necessarily implementing them (Brass and Haggard 2007). A person may decide to act, change their minds, or decide to stop the action they were planning, actions which they claim can really capture what we think of as volition.

According to their research, Brass and Haggard argue that there is an identifiable brain network for these types of voluntary intentional actions, a brain network which is centered on the fronto-median cortex. This network forms intentions about when and how to act, as well as generating the conscious experience of intending (Brass and Haggard, 2007). When people make

intentions that are endogenous rather than a responding to external stimuli, a network of medial frontal areas is more strongly involved. Specifically, this includes the dorsal frontomedian cortex (dMFC) including the pre-SMA and the rostral cingulate zone (RCZ: a region extending from the anterior cingulate cortex and the supplementary and pre-supplementary motor areas), as well as the left and right ventral insula, and the superior temporal sulcus (Lynn, Muhle-Karbe, and Brass 2014, 252).Their research suggests that this network can be partly separated from the brain areas that respond to external stimuli (Haggard and Brass, 2007).

Brass and Haggard's 2007 study on intentional inhibition also specifically considered Libet's proposal of conscious control, where there may be a moment before the action being prepared is completed when we are consciously able to passively let the action happen or more actively veto or cancel the action (Libet, 1985; Brass and Haggard, 2007; Haggard, 2008). When Brass and Haggard (2007) asked participants to perform key presses at a time freely chosen, they also instructed participants to sometimes cancel the action at the last possible moment. In these trials, where the action was cancelled at the last moment, Brass and Haggard found that an area of the anterior frontomedian cortex, rostral to the pre-SMA was activated more than in trials where subjects did not cancel their actions. This may relate to research that has shown that the anterior frontomedian cortex, (BA 10 specifically), may have a crucial role in the storage of conscious action plans (Soon et al. 2008).

4.1.3 Has the Neuroscience of "Volition" Been Captured?

The researchers discussed in this section have used variations on a similar experimental design meant to capture the neural activity that may underlie human volition (Libet et al., 1983; Libet, 1985; Haggard and Eimer, 1999; Brass and Haggard, 2007; Haggard, 2008; Soon et al.,

2008; Lynn et al., 2014). Within the literature of the neuroscience of free will, Libet's work has had a large impact, raising important questions about the way we understand conscious free will. Although his studies have been generally supported by more recent follow-up work, many have raised concerns about Libet's interpretation of his data as well as his conclusion that the results undermine the idea of conscious free will. Some have taken issue with various aspects of the design of the study, the conclusions about free will that Libet has drawn from his studies, as well as more general conceptual concerns (Mele 2007; Roskies 2010; Sinnott-Armstrong and Nadel 2011).

For instance, questions have been raised about whether readiness potential (RP) signals a decision or whether it is better understood as signaling more general preparation for action, or possibly signaling an urge or an intention rather than a decision (Roskies 2010, 2011; Mele 2007, 2011). Others have questioned Libet's reliance on subject's self-reporting the moment they became consciously aware of the urge to move (Banks and Isham 2011). How Libet's studies bear on philosophical and legal concerns about free will and moral responsibility more generally have also been questioned (Sinnott-Armstrong and Nadel, 2011), examining whether we can make sense of moral and legal responsibility without the concept of a conscious will.

Some related key questions I will look at in this section are: how these studies on the neuroscience of free will are defining and using the concept of volition? Can an action such as choosing to press a button or not, really capture human volition? And lastly, do we need to exclude external signals in order to understand the neural correlates of volition?

Considering the first question, what do the various researchers mean by volition and how do they define this concept? For Libet, he specifies that,

> [A]n act is regarded as voluntary and a function of the subject's will when (a) it arises endogenously, not in direct response to an external stimulus or cue; (b)

> there are no externally imposed restrictions or compulsions that directly or immediately control subjects' initiation and performance of the act; and (c) most important, subjects *feel* introspectively that they are performing the act on their own initiative and that they *are free* to start or not to start the act as they wish. (Libet 1985, 529)

Libet's definition of volition is quite similar to Haggard's view of volition. Haggard states that, "volition consists of a series of decisions regarding whether to act, what action to perform, and when to perform it" (Haggard 2008, 934). Although Haggard's definition is less specific than Libet's, we can gather from his research that he also focuses on actions that are deemed endogenous and not guided by external signals, as well as placing importance on whether the subjects *feel* that they are choosing the action, as well as choosing *when* they perform it.

However, a potential issue with the experimental design of these studies concerns the kind of action and decision that is being performed in the experiments. Specifically, can we refer to such a precise and small movement, such as choosing to press a button, or moving one's finger or wrist, as representative of volitional action? Especially given that it is part of a constructed experiment with specific instructions and presumably, none of the action choice options are likely to be preferred or valued in any real sense by the subjects being asked to choose to perform these actions. In other words, the subjects may be choosing when to start or cancel the action, but the choice itself is arbitrary. It would not make a difference for the subject to choose in one way or another.

This issue raises two specific problems. First, the experiments are designed in such a way to study the neural activity involved in the intention to prepare a specific motor activity (including deciding between the action alternatives) all the way through to the completion of the motor act (or in certain trials, the cancellation of this motor act). Setting aside the question of

whether we are in fact looking at a type of voluntary, endogenous decision-making, can we compare a decision between complex, abstract, and value-related questions (e.g., questions about or life choices, or abstract moral conflicts) to a simple motor decision like whether to move our finger or not? There seems to be an important difference between choices that are motivated by our own personal goals and values versus a very simple motor choice, as well as a key difference between motor-related choices and choices that may not involve physical movements.

We can see some parallels between these neuroscientific conceptions of voluntary self-generated action/inaction with many of the philosophical theories of personal autonomy. Importantly, many theories of autonomy emphasize the requirement that an agent's intentions to act are internally generated for an agent to be considered autonomous. This emphasis can also be seen in both Libet's research and Brass and Haggard's research. Both groups have designed their experiments to try and capture that endogenous element of action that seems to capture something causal about the concept of free will that is of central importance. However, the question of whether these experiments can be translated into real-world autonomous decision-making certainly still remains. Despite this criticism, Haggard argues that "these studies do capture a key computational feature of voluntary action, namely that the participant must themselves generate the information that is needed to perform an action" (Haggard 2008, 934). Therefore, although there may not be a perfect translation between these experiments on volition and how we make decisions in the real-world, we may still be able to glean some important information out of them.

This requirement of endogenously generated action relates to the last question I would like to consider—the emphasis in both Libet and Brass and Haggard's experiments on the removal of external signals influencing action choices. Many philosophical theories of personal

autonomy place a great importance on the removal of external influence on an agent's action. That is, the cause of an agent's action must come from within themselves and not be influenced by external people or events. As noted, both Libet's experiments, and Brass and Haggard's experiments, make a determined attempt to remove external signals and to instruct subjects to act according to their own internally generated "urges." Brass and Haggard (2007) contrasted the endogenous action trials with stimulus-driven (or externally signaled) actions in their attempt to identify the specific neural correlates of endogenous decisions to act or cancel actions. The irony that subjects are being instructed by someone external to them to act voluntarily and endogenously is acknowledged by Haggard, but it's worth asking if this emphasis on removing external signals is really necessary. The obvious answer, following the typical conceptions of personal autonomy, is that a focus on internally generated actions necessarily means that a subject's desire or urge to act should not be constrained or influenced by anything external to them. Brass and Haggard seem to distinguish their research on intentional inhibition from the rest of the inhibition or self-control literature, based on their focus of an endogenous voluntary action. They imply that other research (termed proactive inhibition) may be differentiated on the basis of the inclusion of externally-guided events, as compared to the intentional inhibition (or reactive inhibition) which is entirely guided by internal states. A recent article by Lynn, Muhle-Karbe, and Brass (2014) that discussed the difference between the two types of inhibition, they state the following,

> Although proactive inhibition can be considered as resulting from endogenous preparation processes, it still differs from the scope of intentional inhibition as outlined here. Most importantly, proactive inhibition is still concerned with stimulus-guided behaviour in which the outcome (action or omission thereof) will be determined by external events. Whereas intentional inhibition refers to the cancellation of behaviour on the basis of free choice and internal states. (Lynn, Muhle-Karbe, and Brass 2014, 252)

However, some of this proactive inhibition literature on the contrary, claims that proactive inhibitory control is generated by the subject's goals rather than being triggered by external signals (Aron 2011). Meaning, we need a better definition of what is considered external to the agent within this research because it would be odd to claim that a person's goals are external to them. This is especially so, given the fact that goal-related decision-making is primarily driven by internal motivational states.

Additionally, the artificial conditions of these experiments that omit any external elements may detract from us gaining an understanding of what it means to make autonomous (or free) decisions within normal everyday environments when we *do* have a variety of external environmental components impacting our behaviours. Studies that consider how a person makes a decision that they feel is endogenous within an environment with various external, environmental influences may be more important for understanding the neuroscience of personal autonomy or freely-willed decisions, as opposed to studying our decisions when those external signals are already artificially excluded.

Overall, the research on voluntary endogenous actions and intentional inhibition provides some interesting insights to underlying brain circuits involved in these processes. Whether these experiments do in fact capture volition is still debatable. However, they do point to some key questions relevant to many theories of autonomy: do our actions really need to be self-generated and internally caused to be considered autonomous? What ought to be considered internal and external to our decisions? And lastly, do we need to be consciously aware of our intentions before acting for the action to be autonomous?

4.2 The Feeling of Control: Conscious Awareness of Agency and Ownership

In addition to Brass and Haggard's work on voluntary actions and intentional inhibition, many of their studies have also considered the neural correlates of what they have termed "the conscious feeling of agency." There seems to be two distinct subjective feelings underlying this type of conscious awareness: (1) the feeling or awareness of an intention or urge to act, and (2) the causal feeling of ownership over one's actions (e.g., "my body moved" and "I caused it to move") (Haggard 2017). Ideally, we have both a sense of agency about our actions as well as a sense of ownership over our actions. However, some conditions (psychiatric, neurological, as well as some movement disorders) seem to interfere with the processes that produce these conscious feelings. Structural and functional imaging, as well as clinical studies on some of these conditions, have provided some interesting insights into the workings of the networks underlying the sense of agency.

4.2.1 "Autonomy Disturbances" and the Brain

As noted, there has been some interesting research examining the way different conditions or injuries to certain brain regions or networks can impact the way we experience the feeling of agency. For instance, alien hand syndrome (AHS) is a rare disorder involving involuntary movements of a limb (AHS primarily involves the hand, or arm, but can less commonly involve a leg as well) coupled with a sense of loss of ownership over the limb (Hassan and Josephs 2016). Patients with alien hand syndrome often report that their hand seems to have a will of its own—often acting on its own without any conscious awareness. Additionally, some patients adamantly believe that their hand is not their own hand, that it is alien to the rest of their body. Current research seems to agree that there are three main variants

of AHS, each contributing to unique symptoms. Two are particularly relevant to the discussion here: the frontal variant of AHS, and the posterior variant of AHS.

The frontal variant typically stems from a lesion on the left supplementary motor area (SMA) or the left medial prefrontal cortex (mPFC), and can sometimes also involve the anterior cingulate gyrus, anterior corpus callosum, or callosal genu (Hassan and Josephs, 2016). The patients with this variant of AHS are aware that their limb belongs to them but are completely unable to control or suppress its movements (Hassan and Josephs, 2016). That is, their sense of ownership over their body is intact, but their sense of agency is affected.

The posterior variant of AHS, on the other hand, involves "strong feelings of estrangement from the affected limb" (Hassan and Josephs, 2016). In this variant, patients seem to lack a sense of ownership over their limb, reporting that the limb does not belong to them. This type of variant commonly stems from a lesion in the right parietal lobe, and more uncommonly, may involve the occipital lobe and thalamus (Hassan and Josephs, 2016). The research on the different variants of AHS point to important roles of the parietal lobe and the supplementary motor areas in the subjective feelings of body ownership and agency, respectively.

Another related group of disorders are commonly referred to under the umbrella term of environmental dependency disorders—where patients demonstrate various, seemingly non-voluntary actions called "utilization behaviours." These types of environmental dependency disorders are commonly referred to as "disorders in personal autonomy" (Iaccarino et al. 2014, 4). Patients with utilization behaviours automatically (and many times inappropriately) search for and use objects in their immediate environment (Conchiglia et al. 2007, 1). That is, the environment automatically triggers certain actions without their conscious awareness or

intention. The patients are both generally unaware that their hands are inappropriately grasping objects, as well as unable to inhibit the actions. While the frontal variant of AHS typically stems from one lesion on the left SMA, bilateral lesions on the SMA typically result in utilization behaviours (Hassan and Josephs, 2016). Utilization behaviour, in addition to the research on AHS, suggests an important role for the SMA in the inhibition of actions and suppressing irrelevant external environmental cues.

Environmental Dependency Disorder was initially termed by François Lhermitte when describing a particular patient's behaviours following frontal lobe damage (Lhermitte et al. 1986). This patient had a very particular type of loss of control over their environment and identity (Conchiglia et al. 2007, 1). This loss of control led the patient to imitate behaviour around him, apparently without any voluntary intention or purpose, and was unable to inhibit this imitative behaviour (Conchiglia et al. 2007, 1). This imitation syndrome, now sometimes called "Zelig-like syndrome," was famously portrayed in Woody Allen's 1982 movie "Zelig." In the movie, the lead character Zelig is called the "Chameleon Man." His personality and mannerisms entirely change in every social context he is put in, and he seems to have no consistent or coherent identity of his own. For instance, when in the presence of a psychiatrist, he speaks and acts as if he is a psychiatrist, when around a Rabbi, he speaks and acts as if he is a Rabbi, and so on. The character in the movie not only behaviourally adapts to those around him, but he also physically changes as well, hence the nickname "Chameleon Man." In real life, this rare syndrome does quite similarly cause a loss of identity when the patient changes their identity and behaviour according to their environment, but of course without the exaggerated physical changes as portrayed in the movie.

In real life cases, a patient is typically "characterized by a deficiency in personal control over environmental, social and physical stimuli. The patient behaves by responding to stimuli that are implicit in the situation of the social context in which he finds himself" (Conchiglia et al. 2007). According to studies of patients that seem to exhibit these symptoms, just like the movie, the patient will typically take on the characteristics of the people that are in their immediate environmental context.

For example, experiments performed by Conchiglia and colleagues in 2007 on a patient called "AD," found that when being examined by a physician, AD would take on the role of a physician as well, speaking, acting, and believing he is in fact a physician (Conchiglia, Della Rocca, and Grossi 2007, 3). Experiments that observed his behaviour in different environmental contexts found that when they placed him in a bar in the company of the bartender, he completely assumed the role of a bartender. When experimenters observed his behaviour in the kitchen in the presence of a chef, AD assumed the role of a chef. According to Conchiglia et al.,

> [H]e 'constructs' an ad hoc identity corresponding to the context in which he finds himself on the basis of acquired semantic knowledge…AD seems, moreover, to have lost the capacity to keep his own identity constant, as he adapts himself excessively to variations in the social contexts, violating his own identity connotations in order to favour a role which the environment proposes. We might speculate on the existence of a system of specialized monitoring for a form of control that is dedicated to keeping his own identity coherent, independently of the environment. (Conchiglia et al. 2007, 4)

These types of autonomy disturbances can help us piece together a sense of personal control that seems necessary for how we commonly understand the capacity for autonomy to function. There seems to be a link between our feelings of control over our actions and our sense of identity. Whether there is in fact a specific process of control that keeps our identity independent of our environment, as suggested by Conchiglia and colleagues (2007) is still to be

determined. There has been some interesting recent work on the neuroscience of the self indicating that cortical midline structures (e.g., the vmPFC, the perigenueal anterior cingulate cortex (PACC), the posterior cingulate cortex (PCC)), are consistently activated during self-referential processes (Northoff 2017). This network may relate to our sense of self and how this is maintained over time. Although disorders such as alien hand syndrome and environmental dependency disorders such as the Zelig-like disorder are relatively rare, research also suggests that schizophrenia, a more common psychiatric condition, may also impact feelings of agency in addition to other psychiatric diseases that involve psychosis (Haggard 2017). Furthermore, utilization behaviours have also been linked to more common illnesses such as major depression, ADHD, and fronto-temporal dementia (Iaccarino et al. 2014).

According to Haggard (2017), "the key neural correlate of sense of agency might lie in the connectivity between frontal and prefrontal motor areas that initiate action and parietal areas that underlie the monitoring of perceptual events, rather than in any single structure" (Haggard 2017, 204). He cites recent studies on the sense of agency in patients with schizophrenia as evidence in support of this connectivity view (Haggard 2017, 204). For instance, some patients with schizophrenia have symptoms related to the control of their own actions. These patients may feel like their thoughts and actions are not their own, but rather, are caused by external agents (Haggard 2017, 204). To explain this feature, Haggard (2017) cites a model proposed by Frith, Blakemore, and Wolpert (2000), which suggests that these symptoms of control may arise from an inability to predict the consequences of self-generated actions (Haggard 2017). As Haggard (2017) explains, "psychosis thus seemed to abolish the normal prospective element of agency that characterizes healthy adult life…in schizophrenia, the brain may be in a constant state of surprise, attempting to understand the events that it has itself generated" (Haggard 2017,

205). That is, without correct predictions, the sensory experiences that are a result of one's own actions and thoughts are instead perceived as being caused by external events or people.

4.2.2 The Conscious Experience of Intentions and Control over our Actions

Much of the research investigating our sense of agency suggests the crucial involvement of the parietal cortex in the experience of the conscious awareness of an intention to act (Desmurget et al. 2009; Haggard 2017). One study carried out by Desmurget et al. (2009), used direct electrical stimulation (DES) of the parietal and premotor cortex regions of the human cortex to investigate this further. They found that when the right inferior parietal areas were stimulated, the result was a strong intention and desire to move the contralateral limb, whereas stimulation of the left inferior parietal region generated the intention to move the lips and to speak. When the intensity of the stimulation was increased in the parietal regions, Desmurget et al. found that participants believed they had performed these movements, despite the fact that no movement had occurred (Desmurget et al. 2009). In contrast, stimulation of the premotor region triggered overt movements, yet the participants firmly denied that they had moved. Thus, Desmurget et al. suggest that conscious intention and motor awareness arise from increased parietal activation before movement execution (Desmurget et al. 2009).

Interestingly, another study often cited by Haggard, differentiates the conscious feelings associated with the SMA versus the parietal area. In this study by Fried et al. (1991), they showed that intentions triggered by stimulation in the supplementary motor area (SMA) produced the feeling of a compulsion, in contrast to those triggered by stimulation of the parietal area which produced the feeling of voluntary intentions to act (Fried et al. 1991). Another notable experiment recently done by Haggard and colleagues (Caspar et al. 2016), looked at the

relationship between coercion and a sense of agency. Results of this experiment suggested that obeying instructions reduces the sense of agency, "producing an experience that is closer to passive movement than to voluntary action" (Haggard 2017, 206). This is consistent with the commonsense idea that we feel less control over actions that we were coerced into doing, rather than those that are freely chosen without coercive influence. However, an interesting point noted by Haggard is that this research can also help "explain how acting under coercion can lead individuals to experience a psychological distance from unpleasant consequences of their actions…it also demonstrates how social situations can strongly influence an individual's sense of agency, and thus their feelings of responsibility" (Haggard 2017, 206).

Overall, the research on the conscious experience of agency and control over our actions demonstrates some key areas of the brain implicated in these experiences. The various types of autonomy disturbances that can be a result of certain conditions, or damage to certain brain regions or circuits, demonstrate that the way we consciously experience agency has a deep impact on the way we conceive of our autonomy over our actions. This area of research also highlights the link between how we hold a coherent identity constant amongst different environmental factors and our sense of personal autonomy, suggesting that this also may be a key element to the way we consciously experience agency.

Conclusion

In summary, the neuroscientific research on self-generated voluntary action and our conscious feelings of agency highlights some key circuits in the brain that impact autonomy-related processes. Although these lines of research are primarily focused on volition rather than

personal autonomy, they still have some interesting parallels and contribute to an understanding of the way the brain produces behaviours related to autonomy.

When looking at the neuroscience of voluntary actions and intentional inhibition, Haggard and colleagues have pointed out two circuits, the BG-pre-SMA circuit and the parietal-premotor circuit, as central to implementing different kinds of decisions of relevance to voluntary actions (Haggard 2008, 936). Brass and Haggard's study on intentional inhibition also provided some key information related to Libet's hypothesis that there may be a moment before the action being prepared is completed, where we can consciously approve or veto the action (Brass and Haggard 2007). In their trials, where the participant chose to veto the action, Brass and Haggard found an area of the anterior frontomedian cortex, rostral to the pre-SMA was activated more than in trials where subjects did not cancel their actions. They conclude that this area may play an important role in that it is involved in the storage of conscious action plans.

Importantly, Haggard's research points out that it is difficult to pinpoint exactly when and where voluntary actions are initiated. If our circuits for voluntary action are not linear chains that we can continue tracking all the way back to the internal "source of our will" then we may be asking the wrong questions about the causes of our actions. We do know that the basal ganglia integrates a wide range of signals, modulated by dopaminergic input, which coordinates appropriate actions for a particular context. We also know that this process takes into consideration reward or motivational value. This leads Haggard to suggests that voluntary action would be better characterized as flexible, intelligent interaction between us and the current historical, environmental context, rather than an uncaused, "willed" initiation of action (Haggard 2008, 936).

After examining the relevant research, I considered whether Libet's and Haggard's aim of capturing the neuroscience of volition was achieved. I raised some questions about whether the focus on the specific movement of pressing a button could really capture the complexity of volitional action. Relatedly, I argued that the heavy focus of Libet and Haggard on excluding any external signals within their experiments may be too artificial to capture the way we make decisions in everyday life. I suggest that neuroscientific studies looking at the concept of autonomous decisions may instead need to look at how we make decisions within environments that have a lot of external factors to really understand how we manage and control them while still making decisions of our own. Despite potential criticisms of these experiments, they do point to some key questions relevant to many theories of autonomy: Do our actions really need to be self-generated and internally caused to be considered autonomous? What is considered internal and external to our decisions? And do we need to be consciously aware of our intentions before acting for the action to be considered autonomous??

In the second section of this chapter, I turned to the research on the conscious feelings of agency, including the feelings of awareness of an intention or urge to act and the causal feelings of ownership over one's actions (Haggard 2017; 2008; Desmurget and Sirigu 2012; Brass and Haggard 2007; Bayne and Pacherie 2007). Ideally, when acting, we have a conscious feeling of both agency and ownership over our actions. However, I reviewed some research on various conditions where these conscious feelings have been largely impacted, resulting in so-called autonomy disorders. This research highlights some interesting links to the way we experience our identity and are able to hold our identity constant amongst various environmental pressures. Overall, looking at the structural and functional imaging of individuals who have these conditions provides some interesting insights into the way these conscious feelings of agency

and ownership are supported by particular networks in the brain. Yet, Haggard importantly points out that it may be best to understand the neural correlates of the sense of agency as stemming from the connectivity between frontal and prefrontal motor areas responsible for initiating action and the parietal areas of the brain that monitor our perceptual events.

The research discussed throughout this chapter on voluntary action, conscious control over our actions, and our feelings of agency and ownership, will be further explored in the next two chapters where I consider the autonomy theories of Christine Korsgaard and Harry Frankfurt. I will examine whether there are any tensions between what we have seen so far in the neuroscience reviewed and the two prominent accounts of autonomy.

Chapter 5

Korsgaard's Theory of Autonomy: Exploring the Implications of Integrating Cognition, Emotion, and Motivation

Introduction

The previous chapters examined empirical research on various processes relevant to the concept of autonomy, such as decision-making, self-control, voluntary action, and the conscious feelings of agency. Additionally, the research highlighted the key areas where emotion and motivation play a central role in these processes. This chapter will examine how this empirical research bears on Christine Korsgaard's theory of autonomy. I will point out where there are some substantial tensions between Korsgaard's use of concepts such as emotions, reason, and desires, and how these concepts are understood from a neuroscientific perspective. I argue that these tensions indicate some fundamental incompatibilities between the two approaches which raises some important questions about Korsgaard's requirements for autonomy.

First, I will discuss the differences between Korsgaard's use of the term emotion and more recent neuroscientific research on emotions and how emotion, cognition, and motivation are integrated in the brain. Following from this, I will also look at Korsgaard's conception of the self, one that is unified and in control when reason governs emotions. I consider how this conception may differ from a neuroscientific understanding of self-control which views emotion, cognition, and motivation integrating to produce behaviours like decision-making and control over our goal-directed behaviour. Third, I will consider the differences between Korsgaard's use of motivational concepts like incentives and desires and what the empirical literature supports in terms of how emotion and motivation impact control processes. And last, I will consider

Korsgaard's concept of a "reflective distance," and whether this type of conscious reflection of our desires can be supported by neuroscientific research on decision making and motivation processes. Overall, I hope to point out some inconsistencies between Korsgaard's theory of autonomy and what we have seen from the empirical literature, and how these differences might affect Korsgaard's requirements for autonomy.

5.1 Korsgaard's Reasons-Responsive Account of Autonomy and the Role of Emotions

Reasons-responsive theories of autonomy suggest that to be considered an autonomous agent, we must act based on reasons and exercise control over our desires. From this perspective, how we deliberate over our desires and what constitutes rational deliberation, is a central question. Following this tradition, Korsgaard's reasons-responsive account also focuses on our capacity for reason. Reason, for Korsgaard, lies in the particular way that we are self-conscious as humans. We are aware of the fact that we have certain desires or emotions, as well as being aware of our inclinations to act based on those desires or emotions. That is, we are consciously aware of our potential grounds for action (Korsgaard 2009, 115). She claims that this puts us in a unique position, "once we are aware that we are inclined to act in a certain way on the ground of a certain incentive, we are faced with a decision—should we do it?" (Korsgaard 2009, 115). The very fact that we can ask ourselves this question makes it possible for us to be rational, autonomous agents. This specific type of self-consciousness opens up what she calls a "reflective distance" between our motivations and our rational decision to act.

Korsgaard's constitutional model of deliberative action begins with a rational agent that is aware of their desires and the potential grounds for their actions. This awareness, or self-consciousness, gives humans the reflective distance required in order to be in control of these

desires. When the desires are under our control in this sense, we do not need to automatically act on these desires in the way that non-human animals must act on their instincts (Korsgaard 2009, 213). Korsgaard calls the act of deciding whether to act on a desire or not "reason," where you must unify the parts of yourself by making a choice (Korsgaard 2009, 213). Your choice requires a principle, the specified constitutive principles, which ensures that your actions come from your constitutional rule over yourself (Korsgaard 2009, 213).

According to Korsgaard, these constitutive principles come from the categorical and hypothetical imperatives, as well as Plato's principles of a just, well-ordered soul. Deliberation, for Korsgaard, is this act of unifying yourself behind a set of movements in order to count these movements as your own, as movements of a rational agent (Korsgaard 2009, 213). Otherwise, Korsgaard claims, we are "merely a heap of impulses, and not an agent after all" (Korsgaard 2009, 213).

Thus, reasons-responsive accounts of autonomy, as the name suggests, prioritizes the role that reason plays in both decision making and self-control. In particular, to realize one's full agency, we must have the capacity for rational reflection and deliberation. Following from this, emotion is at times seen as an obstacle to our rational decisions which needs to be overcome (Korsgaard 2009, 141). However, Korsgaard's reasons-responsive conception of autonomy does discuss the role of emotions and affective motivational sources in the process of autonomous decision making—although, limiting emotion and desires to very specific roles. This section will discuss Korsgaard's use of the concepts of reason and emotion and compare them to the neuroscientific research on cognition and emotion and how they interact to produce our behaviours.

There are three main areas where she delineates a role for either affective motivation, like desires, or emotion more generally. First, Korsgaard acknowledges the motivational source of action as being affective, or desire-based. Second, she defines emotions as "perceptions of practical reasons," or perceptions of reasons for action (e.g., fear is the perception of a reason to flee) (Korsgaard 2009, 112). From this definition, we are to understand emotions as conscious responses to our practical reasons. And third, she specifies the positioning of emotion in the tripartite soul in a unified autonomous agent—where the emotional or spirited part agrees, along with the appetitive part, to be ruled by reason (Korsgaard 2009, 141). For Korsgaard, when emotion is properly governed by reason, the agent is in control of herself and her actions. Reason then, can guide our actions toward the good of the agent as a whole—whereas emotion would otherwise only be able to direct our actions toward desired objects.

The introduction of emotion into reasons-responsive accounts of autonomy brings reasons-based theories closer to updated understandings of how our brains work when making decisions. There aren't many, if any, reasons-based accounts of autonomy that would entirely deny a role for emotions in autonomous actions today[1]. However, this has also brought new questions into the traditional debate, such as whether acting on emotions can be reasons-responsive in a substantial way (Tappolet 2014, 171). Most reasons-responsive accounts, even those that acknowledge emotions as motivational sources of action, still claim that emotions cannot on their own be responsive to reasons (e.g., Korsgaard's approach is an example of this, also see Tappolet (2014) for full discussion of this point). While there have been other reasons-responsive accounts that have incorporated the idea that emotions can be motivating without

[1] As discussed here, Korsgaard's reasons-responsive account discusses the role of affect as the motivational source of action. See also newer approaches to this subject, such as Karen Jones, "Emotion, Weakness of Will, and the Normative Conception of Agency" (Jones 2003) and Christine Tappolet "Emotions, Reasons, and Autonomy" (Tappolet 2014).

sacrificing our autonomy,[2] Korsgaard still maintains that deliberation (or reason), the capacity to reflect on one's practical reasons, is a necessary requirement for full autonomy.

Despite some of the newer reasons-responsive accounts being sympathetic to emotions having a role in autonomous decision making, there is still a prominent idea that emotion-based decisions and reason-based decisions are distinct and are often in conflict with one another. For instance, the intuition of our rational decisions being overtaken by our feelings can be seen in the classic example of Huck's moral dilemma drawn from Mark Twain's Huckleberry Finn[3]. This example, recently used by Christine Tappolet (2014) in a discussion of emotions in reasons-responsive theories of autonomy, asks us to consider Huck's moral conflict. On the one hand, Huck wants to help his friend Jim escape enslavement. On the other hand, Huck's rational judgment suggests that that Jim should be turned in to the people pursuing him (Tappolet 2014, 172). Huck is left with two conflicting options; one decision, according to Tappolet, would be to act based on his sympathy for his friend, and the other to act based on his deliberative judgment that he should follow the unjust societal rules and turn in Jim. The question Tappolet raises is whether Huck's decision to act on his sympathy for Jim can be considered responsive to reasons, despite the fact that he is "acting on a feeling." For Korsgaard, the answer would be no, deliberative reasoning is required for acting autonomously. Huck would first need to consciously reflect on his desires for his decision to be considered responsive to reasons, and therefore autonomous.

---

[2] See (Tappolet 2014), where Karen Jones's "agential virtue account" is discussed as an example of this type of account; see also Sabine Doring "Seeing What to Do: Affective Perception and Rational Motivation" (Döring 2007).

[3] This example was previously discussed in (Bennett 1974); (McIntyre 1990); (Döring 2007); (Arpaly 2003). Arpaly discusses this example in terms of accidental morally praiseworthy actions and "reverse akrasia", pg. 92. For example, Huck had decided to turn in Jim at the first opportunity he had, but then in that moment was unable to do so, despite his previous deliberation.

However, this example highlights a potential problem in the way we are discussing emotions and reason in decision-making. Despite these approaches bringing emotions into how we understand decision making and trying to make sense of the role emotions play in reasoning, we are still operating within a problematic reason/emotion dichotomy in how we understand decision-making—the presupposition that emotional and rational decisions are different in kind and compete for control over our actions. The empirical evidence reviewed in earlier chapters questions this presupposition, suggesting instead that affective information plays a central role in decision-making. In fact, *all* of our decisions are going to involve emotional and motivational information. Therefore, the moral conflict Huck is facing is not an emotion-based decision competing with a reason-based decision.

The neuroscientific research related to the integration of cognition, emotion, and motivation in a wide variety of processes has also led some researchers to persuasive arguments about the dissolution of mental categories such as cognition, emotion, and motivation (Pessoa, Medina, and Desfilis 2021; Barrett and Satpute 2019; Hoemann and Barrett 2019; Lindquist et al. 2012; Lindquist and Barrett 2012). They argue that the neuroscientific research does not support a difference in kind between cognition and emotion. I will further discuss this point after examining some recent research on the neuroscience of emotions and the implications of this research for Korsgaard's theory of autonomy.

Although there have been huge gains in understanding the neural mechanisms underlying emotions and decision-making, there are many competing theories of emotion that are prevalent from different groups of researchers[4]. A recent, prominent theory of emotion is the psychological

[4] There is a long history of competing ideas and theories about emotions (e.g., evaluative, feelings, motivational, and somatic approaches, for instance). See Andrea Scarantino "The Philosophy of Emotions and Its Impact on Affective Science" (Scarantino 2016) for a detailed discussion of the various theories of emotions.

constructionist view, which aims to provide a neurobiologically plausible theory of emotions (Barrett, 2017; Lindquist et al. 2012). This approach is in line with the integrated cognition-emotion-motivation approaches discussed in previous chapters and is built upon extensive neuroscientific research. Importantly, this view acknowledges that emotions are widely distributed across many areas of the brain. On this view, emotions are seen as emerging from basic psychological operations in the brain, such as core affect, conceptualizations, emotional words, and executive attention (Lindquist et al., 2012). This approach maintains that emotions emerge from basic psychological operations when people make meaning out of the sensory input, both from the body and from the external environment using their own knowledge of their past experiences (Lindquist et al. 2012, 142).

The psychological constructionist approach[5] to emotion is in line with other cognitive neuroscientific frameworks which support the argument that neural networks are not necessarily highly specialized to certain functions or tasks, and in fact, are often activated across different task domains (Pessoa 2022; Shine and Poldrack 2018; Anderson 2014; Anderson and Pessoa 2011; Poldrack et al. 2009; Price and Friston 2005). In other words, from this perspective, large-scale, neural networks are activated across many different categories of tasks (e.g., emotional, cognitive, perceptual) and interact to produce the psychological event or behaviour (Lindquist et al. 2012). The research from this approach has shown that the experience of emotion increases activations in other brain regions, such as,

> [A]ctivation in brain regions implicated in conceptualization (simulation of prior episodic experiences), language (representation and retrieval of semantic concepts), and executive attention (volitional attention and working memory),

[5] See also Schröder and Thagard, "Emotions as semantic pointers: Constructive neural mechanisms" (Schröder and Thagard 2015) for a detailed discussion of their neurocomputational theory of emotions that is compatible with the psychological constructionist approach to emotions.

> suggesting that these more "cognitive" functions play a role in constructing experiences and perceptions of emotion. (Lindquist et al. 2012, 140)

From this perspective, people automatically make meaning out of their core affective state[6] in a particular context and experience it through categorization as "part of an emotion, perception, belief, or judgment" (Lindquist et al. 2012, 124). The role of executive attention is to help direct this combination of psychological operations into producing an emotional gestalt by "processing information from the body (core affect), information from outside the body (exteroceptive sensory information), and representations of prior experiences (conceptualizations)… executive attention helps determine which representations are utilized to make meaning of that state, and which are suppressed" (Lindquist 2012, 126).

On this view, experiencing an emotion like fear, for example, can be understood as an embodied, dynamic representation that is highly dependent on a particular context (termed a "situated conceptualization" by Lindquist et al., 2012). Being highly influenced by context helps explain why an emotion can be experienced differently in different situations. For example, an instance of fear can be experienced very differently according to context in which we find ourselves. We might feel fear in the context of being on a rollercoaster at an amusement park or feel fear walking alone down a dark alley at night. While there may be some overlap in features of both instances (e.g., a racing heart), there would also be different context-related features including affective, physiological, action or movement, and brain related features (e.g., feeling excitement for the rollercoaster, but not while walking alone at night) (Fridman et al. 2019, 3).

[6] Core affect is defined by Lindquist et al. (2012) as usually accompanied by somatovisceral, kinesthetic, proprioceptive, and neurochemical fluctuations within the body and are represented in the brain. Changes felt in the core affect are understood to be the body's way of representing when the objects in the environment are valuable or not in a particular context (Lindquist et al. 2012, 124).

This conception of emotions also explains why emotions can be experienced differently, not only within the individual according to context, but also more widely across different people and different cultural and environmental contexts (Barrett 2017a; 2017b).

How would the psychological constructionist view of emotions line up with Korsgaard's conception of emotion? I would argue that this neurobiological conception of emotions is quite fundamentally incompatible with Korsgaard's view of emotions. First, Korsgaard defines emotions as perceptions of practical reasons. According to the psychological constructionist approach, "emotions are constructions of the world, not reactions to it" (Barrett 2017b, 16). Furthermore, the very question of whether we can "act on our feelings," or have reasons for our emotions, implies a linear, causal competition between emotions and reason which is not backed by empirical evidence. Rather, from an integrated perspective, in an important sense, we are always acting on our feelings because reason and emotion are not separate categories. Cognitive, emotional, and motivational information are all contributing to various tasks across traditional domains. For example, emotional information is not only processed in what would be considered an emotional domain, such as when experiencing fear for example. Emotional information also contributes to tasks that are usually considered cognitive or motivational, such as using our memory or the reward system (Beam et al. 2021). This type of integrated cognitive-emotional-motivational perspective posits the relationship between emotion and cognition as cooperative sources of information that contribute to various neural networks from which our behaviours emerge. As discussed by Lindquist et al. (2012),

> We might not assume that emotion and cognition battle it out in the brain when a person makes the moral decision to sacrifice one life to save many…or that consumer decisions are predicated on competing affective and rational representations… Instead, we might assume that affect and executive attention are merely different sources of attention in the brain rather than processes that differ in kind. (Lindquist et al. 2012, 29)

Considering Huck's moral dilemma again, it is presented as a conflict between his emotional decision and his rational decision. However, the empirical literature suggests that this is a problematic view. Both decision options for Huck, in that particular moment, are coming out of the same neural decision-making context. This particular context involves varied information related to his circumstances in that moment (e.g., sensory, interoceptive, affective, environmental factors, personal history, learned behaviours, reward information, predictions for future actions and rewards, any uncertainties that are present, etc.). Huck needs to decide what to do when faced with his moral dilemma. However, his decision options will have the same neural contextual factors. Should he decide on either of those options, both decisions will involve cognitive, emotional, and motivational information. Therefore, even though it intuitively might seem like he is deciding between an emotional decision and a rational decision, he is in fact just deciding on a course of action in that moment. To further explain this idea of a neural context, we can consider Hoemann and Barrett's (2019) recent discussion on this topic:

> Every categorisation of sensation (e.g., as fear) updates the neural context in which the brain is making predictions for the body. The brain will subsequently prioritise perceptions, actions, emotions, and cognitions that have previously been reinforced in similar situations. This iterative process of constructing and confirming predictions gives rise to the folk intuition that cognition and emotion cause one another or compete for control…Rather than having separate causes, cognitions and emotions are constructed based on the temporal dynamics of the brain (Spivey 2007). As the brain transitions through all possible patterns of neural activation (i.e., state space), the current brain state, in combination with inputs from the body and world, influences the probability of future brain states (Barrett 2009). (Hoemann and Barrett 2019, 68-69)

From this perspective, Huck's various decision options may compete in terms of his mental resources, motivational value, and attention[7]. However, it is not a matter of an "emotional decision" competing with a "rational decision," as it has been discussed in relation to the reasons-responsive accounts of autonomy. As the quote above states, cognitions, emotion, motivation, and perceptions all impact the way we interpret the world, as well as how we decide to act in every instance. If these concepts are not considered separately competing processes, we may need to rethink the way these concepts are used in traditional debates about reason and emotion within the philosophical autonomy literature.

5.2. Self-Constitution in Korsgaard's Theory of Autonomy

Relatedly, the integration of cognition and emotion also has implications for Korsgaard's conception of the self that underlies her model of autonomy. Korsgaard's account requires a type of self-control characterized as a "unified self," where reason controls one's emotions. This section will consider Korsgaard's idea of the self, self-control as "reason controlling emotion," and whether this conception of the self can be supported by the neuroscience of self-control.

One of the key aspects of Korsgaard's account is that autonomous decisions must come from an agent with a unified identity. Without this unified identity, we cannot consider the person to be an agent, and by extension, their actions cannot be considered autonomous. Korsgaard claims that our incentives come from our various practical identities that are a reflection of the many roles and relationships we have in our lives. But these various practical

[7] The mechanism for this kind of competition is thought to be parallel constraint satisfaction (Lindquist et al. 2012); see also Pessoa (2013) for a detailed discussion of resource-capacity competition in relation to emotion and cognitive control (Pessoa 2013, 124).

identities must be unified to decide upon which incentive we are to act. For Korsgaard, this unification of the self is the process of deliberation. As she says,

> When you deliberate, when you determine your own causality, it's as if there is something over and above all of your incentives, something which is you, and which chooses which incentive to act on. This means that when you determine yourself to be the cause of the movements which constitute your action, you must identify yourself with your action, you must identify yourself with the principle of choice on which you act. (Korsgaard 2009, 72)

The unification process for Korsgaard is Platonic and requires the rational part of us democratically ruling over the spirited and appetitive parts in order to act for the good of the whole person (Korsgaard 2009, 146). Once unified, an agent is able to decide how they want to act, based on the rational principle of their choosing. The active choosing of a rational principle is Korsgaard's required "agent contribution," which makes their actions autonomous and truly their own. Korsgaard claims that this type of agent contribution is necessary to distinguish between self-determination of one's actions and merely being caused, similar to what a non-human animal may experience (Korsgaard 2009, 107).

There are some significant tensions between Korsgaard's discussion of the unified rational self and the empirical research on self-control. In particular, the key idea behind Korsgaard's account of autonomy is a very specific way that reason controls emotion. This type of self-control is a requirement for full agency and an indicator of agential authority. However, Korsgaard's concept of self-control as "reason controlling emotion" may not be supported by the neuroscience research on self-control processes. In particular, as discussed in Chapter Three, emotional and motivational information has been shown to play an important role in self-control, which may be problematic for Korsgaard's requirements for personhood.

Korsgaard's concept of autonomy requires a particular self-constitution that the agent must have to be able to make decisions. Self-constitution, in her model, requires that the agent's rational part of the self governs the emotional part of the self. She distinguishes her constitutional model of the self from what she refers to as the "combat model," where reason is pitted against the passions and reason tries to forcibly repress the emotional parts of the self (Korsgaard 2009, 133). Rather, for Korsgaard, the constitutional model, following the Platonic model of the soul, sees the interaction between reason and emotion as analogous to a well-run city, where the people democratically agree to be ruled for their own benefit. A united self, similarly, would agree that reason should rule over the appetitive and spirited parts for the good of the agent (Korsgaard 2009, 135).

With Korsgaard's model of deliberative action being influenced both by Plato and Kant, she further suggests that Kant would also view the interaction between reason and emotion as constitutional rather than combative (Korsgaard 2009, 152). "For Kant, like Plato, reason must rule for the good of the whole, and if we identify with the voice of reason, it is only because we identify with our constitution, and it says reason should rule" (Korsgaard 2009, 158). Korsgaard's conception of the self in her constitutional model of deliberative action then can be understood as analogous to Plato's tripartite soul, with reason governing the lower parts of the soul. She makes the analogy as follows,

> Deliberative action begins from the fact that we have certain appetites and desires. We are conscious of these, and they invite us to do certain actions or seek certain ends. Since we are self-conscious however, we do not act on our appetites and desires automatically, but instead decide whether to satisfy them or not. And then finally there is carrying the decision out—actually doing what we have decided to do. For of course we don't always do what we have decided to do, but are sometimes distracted by temptations, pleasure, pain, and fear from the course we have set for ourselves. So we can identify three parts of a deliberative action corresponding to Plato's three parts of the soul, namely:

> Appetite makes a proposal, reason decides whether to act on it or not, spirit carries the decision of reason out. (Korsgaard 2009, 141)

When our self is unified in this way, an agent's movements and decisions can be attributed to herself as autonomous. The agent would not be identifying with reason specifically, but rather identifying with their own constitution. This is crucial for attributing agential authority in Korsgaard's model, where the self is united as something above and beyond its parts to have the capacity to self-govern (Korsgaard 2009, 141). However, this also means that her conception of agential authority is directly tied to the idea of reason ruling over one's emotions. Even if Korsgaard's conception of reason ruling over emotion is democratic, rather than combative, the Platonic idea of reason being in control of the appetitive and spirited parts of the soul is incompatible with the empirical evidence suggesting that cognition, emotion, and motivation are integrated in processes like decision-making and self-control.

As discussed in the previous section, there is quite a bit of neuroscientific research that supports the idea that cognition, emotion, motivation, and perception emerge from interactions between domain-general, distributed networks in the brain. As summarized by Barrett and Satpute (2013) in a recent overview of this perspective, they state that,

> Over two decades of brain imaging data point towards a framework where the human brain is intrinsically organized into domain-general, distributed functional networks. Emotions, social cognitions, and non-social cognitions (and perceptions…) can be thought of as mental events… that are constructed from interactions within and between these networks that compute domain-general functions. (Barrett and Satpute 2013, 361)

This research suggests that we move away from considering emotions and cognition as separate processes and the intuition that emotion and cognition can "cause each other" or compete for control, as discussed in the previous section. Recent research shows that affective

networks are activated across many different tasks, including those traditionally thought of as strictly cognitive or perceptual. As stated in a discussion on this topic, Pessoa (2013) argues that the empirical evidence supports an integrated perspective on emotion and cognition:

> [C]ognitive-emotional interactions take many forms. Notably, lateral prefrontal cortex, which has well-documented roles in cognition, is also an important convergence site for cognitive and emotional signal interactions. Thus, in many cases, the "cognitive" or "emotional" nature of mental processes is blurred in a way that highlights the integration of these domains in the brain. Thus, too, it is both conceptually and empirically unproductive to carve the brain into cognitive and emotional realms. (Pessoa 2013, 251)

The idea that the lines between cognitive and emotional processes is blurred is well-supported in research from many different research areas. For example, as shown in a meta-analytic review by Lindquist et al. (2012), "emotion" can be found in other task domains such as,

> [A]ction simulation and perception (Grèzes and Decety 2001); autobiographical memory (Svoboda et al. 2006); decision making (Krain et al. 2006); executive control (Owen et al. 2005; Wager and Smith 2003; Wager et al. 2004); language (Vigneau et al. 2006); self-referential processing (Northoff 2006). (Lindquist et al. 2012, 139)

This research suggests that emotion is involved in constituting other cognitive and perceptual events and that, as put by Lindquist et al., separate "categories such as emotion, cognition, and perception are not respected by the brain" (Lindquist et al. 2012, 139).

In addition to the point that an integrated cognitive-emotional-motivational approach does not see cognition and emotion as different-in-kind processes, this approach also suggests a mutually cooperative, rather than combative relationship between cognition and emotion. This framework suggests we view emotion and cognition as cooperative and integrated, where "emotion does not influence cognition during decision making as one pool ball exerts influence on another. Instead, the view suggests that core affect, conceptualization, and executive attention

(and perhaps other psychological operations) cooperate to realize a behavioral outcome" (Lindquist et al. 2012, 142).

Moreover, this view of emotion and cognition suggests that "cognition does not control emotion in a top-down fashion, nor do emotions provoke cognitions; the transition from one to the other occurs in an uninterrupted, domain-general meaning-making process" (Hoemann and Barrett 2019, 69). Therefore, we can understand the relationship between cognition and emotion as cooperative in the sense that they are both integrated and contribute to our mental events or behaviours. However, it is a complex relationship, where emotional information influences what we traditionally think of as "cognitive tasks" and vice versa.

The exact way that emotion and cognition interact in the prefrontal cortex is complex and varied, according to many different factors and the specific context and tasks being undertaken. However, the integrated network framework raises questions about the way we think of self-control, and the roles played by emotion, cognition, and motivation. As discussed in detail in Chapter Three, emotion and motivation have been shown to impact executive control functions in different ways (Pessoa 2018; Yee and Braver 2018). In particular, recent studies have suggested that both emotion and motivation can either enhance or impair our behaviours, depending on how they interact with executive control functions in a specific context (Pessoa 2018, 160). In fact, there have been some studies that show that cognitive tasks can at times counteract emotional information, which would go against the intuition prevalent in the autonomy literature that emotion is more powerful than reason and can override our judgments (Pessoa 2013, 251). This would mean that our intuitions are likely leading us to oversimplify the way emotion and reason interact to produce our behaviours, and the way we understand self-control.

Furthermore, some of the newer neuroscience research on self-control discussed in Chapter Three (Inzlicht, Bartholow, and Hirsh 2015), suggested that emotional information may have a causal role in self-control processes. In particular, the affective-control frameworks suggest that affective signals act as a catalyst signaling that a goal conflict is occurring. That is, the conflict is experienced as aversive or unpleasant, leading to avoidant behaviour (Okon-Singer et al. 2015; Botvinick 2007; Kool et al. 2010 ; Dreisbach and Fischer 2015; Shenhav and Buckner 2014; Dignath et al. 2020; Inzlicht, Bartholow, and Hirsh 2015). As discussed in detail in Chapter Three, this suggests that emotion and motivation are crucial and necessary for self-control. Instead of emotion and reason competing for control—as required by Korsgaard's Platonic model of self-control—reason, emotion, and motivation are all integral and work together to both produce and control our behaviours.

5.3. The Role of Desires and the Motivation Process

For Korsgaard, there are three main types of motivational concepts underlying her theory of autonomy: desires, incentives, and inclinations. Although she does not directly define desires, she at times uses desires interchangeably with the term incentives. Both concepts seem to play the role of presenting possible motivational grounds for action. However, with both our desires and incentives, we are supposed to be able to step back and consciously reflect on whether we ought to act on the particular desire or incentive that is being proposed. For example, Korsgaard says the following on the role of our desires:

> When you are aware that you are tempted, say, to do a certain action because you are experiencing a certain desire, you can step back from that connection and reflect on it. You can ask yourself whether you should do that action because of that desire, or because of the features that make it desirable...and if you decide that you should not, then you can refrain. (Korsgaard 2009, 19)

In this quote, Korsgaard has desires presenting a possible option for an action. Following this, we are supposed to step back from the desire and reflect on whether we ought to perform the action or not. However, later when discussing deliberative actions, she replaces the term desires with incentives[8]:

> So here you are with your free will, completely self-governing, with nothing outside giving you any laws. And along comes an incentive, let us say, a representation of a certain object as pleasant. Being aware of the workings of that incentive upon you, you have an inclination for the object. And that inclination takes the form of a proposal. So the inclination says: end-E would be very pleasant. So how about end-E? Doesn't that seem like an end worth pursuing? Now what the will chooses is, strictly speaking, actions, so before the proposal is complete, we need to make a proposal for action. Instrumental reasoning determines that you could produce end-E by doing act-A. So the proposal is: That you should do act-A in order to produce this very pleasant end-E. (Korsgaard 2009, 153)

The above quotes demonstrate the role that the motivational concepts play in Korsgaard's model of deliberative action. For Korsgaard, "…incentives are…features of the represented object that make it, from some point of view, attractive or aversive" (Korsgaard 2009, 120). Because of our ability to be consciously aware of these incentives, we can then respond to this incentive, therefore transforming our incentives to inclinations. It is then the inclinations that form the proposal of action that we must rationally decide to act on or not (e.g., by rationally assessing if the proposal passes the categorical imperative). While she specifies the source of incentives as coming from our practical identities, she also claims that our incentives (as well as our desires) are still tied to our instincts (Korsgaard 2009, 111). Whereas our inclinations, like

[8] Korsgaard specifies in a footnote that she does not consider incentives and desires to be identical concepts. However, throughout her model she does seem to use them interchangeably. She admits that in previous work she has described both desires and inclinations as incentives, but states that this was incorrect (Korsgaard 2009, 121).

reason, come from our uniquely human self-consciousness. That is, our ability to step back and reflect on our incentives, makes our inclinations possible.

Korsgaard's use of the concepts of incentives and desires as the motivational grounds for action has some similarities to the empirical research on our motivational processes. There has been a lot of research on the idea of incentive salience, or the motivational weight of rewards guiding our decision-making and goal-directed actions. However, there are still some key differences that stand out between the empirical research on the reward network and the motivation literature and Korsgaard's discussion of motivational concepts such as desire, incentives, and inclinations.

Much of Chapter Two was devoted to explaining our motivational system from a neurobiological perspective, to understand the neural mechanisms underlying desires and how and why we decide to act in certain ways. The underlying motivational processes of our action choices or our desires to act or attain something desirable, are often driven by the reward network. The reward network relies on the cortico-basal ganglia circuit (or BG circuit), which importantly connects the nucleus accumbens (NAc) (part of the ventral striatum), the orbitofrontal cortex (OFC), and the anterior cingulate cortex (ACC), three main areas which work to execute motivated and planned actions (Haber 2011). The circuits in the reward network respond to and are coordinated through the action of the neurotransmitter dopamine (Montague et al. 2004). Dopamine released from the midbrain (VTA or SN) is then projected widely throughout the brain, influencing the BG circuits involved in reward-related learning, action, and cognition (Haber, 2011).

There has been a lot of empirical support behind the idea that an important function of dopamine within the cortical-BG-DA circuit, is that dopamine works as an error prediction

signal, helping us learn to make choices based on the probability of positive outcomes or rewards (Frank and Claus 2006; Schultz 2006). From this perspective, learned (and often unconscious) environmental cues help us predict the possibility of rewards and become "imbued with incentive properties" (Hyman 2010, 262). The brain assigns this sort of reward-driven motivational weight to the stimuli, guiding our actions towards the satisfaction of the rewards or goals it is signaling (Mannella et al. 2013). The assigned motivational weight to stimuli, or incentive salience, as it is commonly referred, is thought to activate feelings of "wanting," thereby initiating the necessary steps to obtain the wanted reward or goal (Hyman, 2010). Dopamine is thought to both learn and predict future rewards, but also to bias our attention and action choices (McClure, Daw, and Montague 2003).

This understanding of incentive salience at first glance is similar to Korsgaard's concept of incentives. With incentive salience, the identification of an object, features of an object, or some state of affairs gets assigned motivational value due to its connection to a reward that we want. Korsgaard's definition of incentives is "features of the represented object that make it, from some point of view, attractive or aversive" (Korsgaard 2009, 120). However, despite this similarity, the role that incentive salience versus Korsgaard's idea of motivational incentives play in guiding behaviour, is quite different.

Incentive salience seems to do more direct action-guiding work than Korsgaard's conception of incentives. Incentive salience, or the assignment of motivational weight to an object, is a major factor in the decision to act and pursue the rewards. Whereas for Korsgaard, any decisional aspects do not yet come into the picture during the incentive stage. It is not until incentives are transformed into inclinations that we begin to deliberate on whether we should act on them, in conjunction with our reasoned principles. From Korsgaard's perspective, the

affective motivational elements (as well as emotion), are entirely removed from the deliberation stage of decision-making.

From the empirical perspective on motivational processes, incentive salience can trigger actions that are acted on automatically and effortlessly, without any conscious deliberation (Aarts and Dijksterhuis 2000). These non-conscious deliberative actions help us pursue our goals and any objects or states of affairs that we desire. Cognition, emotion, motivation, and perception are all integrated to guide these goal-driven actions, often without a need for the type of conscious, rational deliberation as required by Korsgaard.

If desires (or rewards) do have a more direct role in determining our actions, or at least in directly motivating certain actions, this would create an issue for Korsgaard's requirement for agential authority. As previously discussed, Korsgaard states that desires cannot directly *cause* our autonomous actions. It is only through the deliberation process, which includes our agent contribution of a rational principle, that autonomous actions can be caused by the agent (Korsgaard 2009, 105). For Korsgaard, this is the difference between self-determination and being causally acted upon. If one does not choose the rational principles used in order to determine whether to act on the proposed incentive, then the agent is not acting from an autonomous place, or from their autonomous self (Korsgaard 2009, 107). Therefore, if rewards directly influence behaviour choices, thereby removing the need for rational deliberation or the agent contribution itself, Korsgaard will need another way to determine agential authority and control over the desire.

5.4 Reflective Distance: Are we Consciously Aware of our Desires?

The way Korsgaard defines the specific "deliberative reason" requirement in her account of autonomy also highlights a possible incompatibility between her theory of autonomy and the empirical literature. In particular, the idea that we must have the capacity to consciously reflect on our desires for autonomy is substantially different from the empirical research on decision-making reviewed in previous chapters. This section will focus on Korsgaard's idea of conscious deliberation and whether the neuroscience research can support this requirement for autonomy.

Within Korsgaard's model of deliberative action, the most important step in acting autonomously is when our incentives are turned into inclinations. When we become consciously aware of our incentives, this opens up what she calls a "reflective distance," or the space that makes reasoning possible (Korsgaard 2009, 119). For Korsgaard, without the ability to be consciously aware of our incentives, we are not able to control them, nor would we be able to act autonomously. Our conscious awareness changes how we experience incentives. Rather than the incentives acting as an uncontrolled force within us (e.g., like instincts are experienced in non-human animals), our ability to consciously deliberate gives us the ability to reflect on and control our incentives (Korsgaard 2009, 119). It is within this space that we are able to apply reason and combine our response to the incentive with a principle of our own choosing, thereby deciding for ourselves how we want to act.

However, are we consciously aware of our desires or intentions to act? The empirical literature discussed in Chapter Four raises many interesting questions on the conscious awareness of our decisions. For example, the literature following the Libet experiments (1982), considers whether there may be a brief moment of conscious awareness in our otherwise unconscious processes of voluntarily deciding to act. Brass and Haggard's (2007) study on

intentional inhibition also specifically considers Libet's proposal of voluntary conscious control over our action choices, where there may be a moment before the action being prepared is completed when we are consciously able to passively let the action happen or more actively veto or cancel the action (Libet, 1985; Brass and Haggard, 2007; Haggard, 2008).

We can compare these two similar ideas—Libet's conscious moment to allow/veto an action with Korsgaard's concept of "conscious reflective space," where our desires present a "proposal" for action that one must consciously reflect on and rationally decide to act. Korsgaard says the following on our awareness of our desires:

> When you are aware that you are tempted, say, to do a certain action because you are experiencing a certain desire, you can step back from that connection and reflect on it. You can ask yourself whether you should do that action because of that desire, or because of the features that make it desirable...and if you decide that you should not, then you can refrain. (Korsgaard 2009, 19)

This idea of a conscious reflective space, as presented by Korsgaard, is similar to Libet's proposal. If, as Libet claims, there is a moment of conscious awareness where we can either permit or veto an action plan, this could be in line with Korsgaard's "conscious reflective space." However, Libet's claim seems to be focused on the conscious awareness of our action decision, rather than the conscious awareness of the initial motivational desire, or the motivational source itself that may have prompted the action. It is this initial desire, specifically, on which Korsgaard claims we must consciously reflect.

This means that while Libet's and Korsgaard's conscious moments may look similar on the surface, they are not referring to the same moment in the action process. Korsgaard requires an awareness of the earliest moment, where we evaluate the motivational source of our actions, not the last step in the action process where we may be able to allow or stop the decided upon action from going through. While we may be able to feel some of our motivational urges at

certain points in the process, we are not likely consciously aware of them so early in the process in the sense that Korsgaard's account requires. Korsgaard claims that in this early conscious moment we would need to apply our choice of rational principles, therefore deciding the best action for ourselves. However, the studies by Haggard and Brass (Brass and Haggard, 2007; Haggard, 2008) as well as the earlier Libet studies (Libet et al., 1983; Libet, 1985), on the other hand, suggest that our brain has already started to prepare our action response. At best, there may be a conscious moment where we can allow this action to happen or cancel the action.

As shown in detail throughout Chapter Two, the way our motivational processes work in order to guide our behaviour effectively are very different from Korsgaard's conception of how our actions are motivated. The motivational/reward network, and then our action planning processes are quite effective at learning what rewards we want, guiding us toward the necessary actions in order to attain these rewards, often automatically and effortlessly without the need for conscious deliberation. The processes that support our autonomous, goal-directed behaviours are not often consciously deliberative in the sense that Korsgaard's account would require for full autonomy.

There does seem to be an underlying intuition within the autonomy literature that consciously deliberating feels qualitatively different than unconscious, automatic actions, and that there is a fundamental difference between these two types of actions which is of central importance to autonomy. This intuition is similar to the debate over the automatic and controlled processes in the neuroscience literature. Controlled processes are considered to require more active attention, and are therefore more effortful, reflective, and slow. They are also considered to use up more of our resource capacity in our brains (Pessoa 2013, 248). On the other hand,

these capacity-limited controlled processes are thought to be balanced out by automatic processes, defined as,

> [A] sequence of nodes that nearly always becomes active in response to particular input configuration, where the inputs may be externally or internally generated and include the general situational context, and where the sequence is activated without the necessity of active control or attention by the subject…certain sensations and mental operations seem to occur "automatically," with no effort. (Pessoa 2013, 248)

The two types of processes, controlled and automatic, have been historically characterized as two opposing systems in the brain. As examined earlier in Chapter Two and Three, this automatic/controlled dichotomy is often invoked in dual process frameworks for decision-making, and commonly referred to as "system 1" and "system 2," where system 1 is considered automatic/heuristic/reflexive and system 2 is considered controlled/analytic/reflective (Pessoa 2013, 250: see also Keren and Schul (2009) for a detailed discussion of the viability of the automatic/controlled distinction in decision-making).

However, recent empirical evidence does not support this strong distinction between autonomic and controlled processes, instead favouring a gradual, or continuous framework. As argued by Pessoa (2013), one of the researchers in favour of the gradual approach,

> Because automatic and controlled processes appear to be qualitatively opposed, it is natural to dichotomize mental phenomena into these two classes. But such a dichotomy has simply not held up in the face of new data: reports of automaticity have invariably been countered by reports of capacity limitation; behavioral effects assumed to operate automatically are affected in ways that belie that assumption. (Pessoa 2013, 249)

I would argue that there are traces of this problem in the autonomy literature as well. It can be seen in the intuition that conscious reflection is somehow more autonomous than behaviours or choices that are made automatically and unconsciously. For Korsgaard's account

of autonomy, this distinction is present in her requirement of conscious, reflective deliberation of our desires, where our desires are seen as automatic (by being tied to our instincts) and impulsive, and therefore in need of rational deliberation to be considered autonomous. The need to prioritize conscious deliberation, as controlled autonomous decision-making, is not in line with the newer models of how our brain makes decisions to act. Rather than a strict dichotomy between automatic and controlled behaviours, it has been suggested that we view these processes as a gradual continuum. The gradual continuum between automatic and controlled behaviours ought to be reflected in the way we discuss these differences within the autonomy literature as well. But importantly, although certain actions may fall higher or lower on the automatic/controlled continuum, this does not reflect how autonomous the decision might be, nor does it tell us something important about autonomous actions. Slow, controlled, reflection or not, an action can be autonomous. Therefore, overall, I would argue that Korsgaard's reflective space as a requirement for autonomy, a space in which reason can be applied to monitor and control our affective motivational grounds for action, is not well supported by the neuroscience of decision-making.

Conclusion

The aim of this chapter was, given the neuroscience reviewed in previous chapters, to assess whether Korsgaard's requirements for autonomy can be supported by this research. I argued that the neuroscience suggests that cognitive, emotional, and motivational information are often integrated in processes such as decision-making and self-control. If this is the case, I argued that Korsgaard's conception of emotion and the role that it plays in autonomous decision-making and control may not be compatible with the neuroscientific research.

Following from the discussion about the integration of cognitive, emotional, and motivational information, I also considered some of the possible implications that an integrated cognitive-emotional-motivational approach might have on Korsgaard's conception of the self. I argued that there are substantial tensions between Korsgaard's account of self-control as reason controlling emotion, as compared with the neuroscience of self-control. If concepts such as emotion, cognition, motivation, are integrated and not separate mental categories, it is unclear if we can make sense of reason controlling our emotions, creating a possible issue for Korsgaard's way of determining agential authority over our actions.

I also argued that the neuroscience research on our motivational reward system may not support the way Korsgaard defines motivational concepts like desires, incentives, and inclinations. Korsgaard's requirement to consciously reflect on our motivational source of action may not be plausible, given that our motivational system often effortlessly and unconsciously guides our actions.

Lastly, I continued this argument by examining more closely Korsgaard's idea of "reflective distance" and the idea of being consciously aware of our desires. I argued that there is an underlying intuition that there is an important difference between "automatic," nondeliberative actions, and "controlled," reflective, and deliberative actions both within the autonomy literature and the neuroscience literature. I suggest that, like the neuroscience debate about automatic/controlled processes, viewing this distinction as a gradual continuum, rather than two competing processes, may be a better approach. Importantly, I argue that the distinction between automatic and controlled behaviours does not, in fact, distinguish autonomous actions from nonautonomous actions.

The newer neuroscience frameworks that support an integrated approach to emotion, cognition, and motivation (Pessoa, Medina, and Desfilis 2021; Barrett and Satpute 2019; Hoemann and Feldman Barrett 2019; Pessoa 2022; 2013; Barrett 2017a; 2017b; Lindquist et al. 2012; Lindquist and Barrett 2012; Shackman et al. 2011) have questioned the validity of the deeply entrenched dichotomization of reason and emotion, where they are pitted against each other, fighting for control over our decisions and control over our self. When we consider cognition, emotion, and motivation as integrated functions of domain-general large-scale networks, some of these dichotomies begin to conceptually break down. I argue that this can be seen to have some direct implications for reasons-responsive theories of autonomy, like Korsgaard's theory and her requirements for autonomy.

Chapter 6

Frankfurt's Theory of Autonomy: Emotions, Desires, and the Experience of Agency

Introduction

The last few decades of scientific research have dramatically impacted the way we understand decision-making, self-control, and motivation. In particular, there is considerable evidence supporting the fact that affect plays a central role in these processes. Within the philosophical autonomy literature, various approaches have also begun to recognize the role of emotions in autonomy. As discussed in the previous chapter, even highly rational, reasons-responsive accounts of autonomy at least acknowledge emotions as potential sources for motivation. Frankfurt's theory of personal autonomy is focused on the motivational aspects of an agent's decisions, in particular, the way desires can move us to act. Thus, this approach is usually thought to be more affective by nature, particularly compared to reasons-responsive accounts of autonomy. As such, his approach stands out as one that could potentially be more compatible with an up-to-date understanding of integrated cognitive-affective processes in the brain.

This chapter will examine whether Frankfurt's account of personal autonomy is in fact compatible with the neuroscience of decision-making, self-control, and the conscious feelings of agency reviewed in earlier chapters. I will examine some of the key ideas in both Frankfurt's earlier hierarchical approach and his more recent caring approach to autonomy, comparing these ideas with the empirical research. I first consider the implications for Frankfurt's theory considering the role that affect plays in the neuroscience of decision-making. Second, I examine the way Frankfurt conceives of desires operating within the motivational structure of an agent. I

discuss possible tensions between this conception and the research on the neuroscience of goal-directed behaviours. Third, I look at Frankfurt's unwilling drug addict example and compare this with some recent research on the neuroscience of addiction and the ways in which addiction impact decision-making and self-control processes. Lastly, I look at the neuroscientific research on the conscious experience of agency and control and consider how this may impact Frankfurt's requirements for agential authority over our actions. Throughout this chapter, I point out some key incompatibilities between Frankfurt's approach to autonomy and the recent research on decision-making, self-control, and our motivational system.

6.1 Reason and Emotion in Frankfurt's Theory of Autonomy

Frankfurt's theory of personal autonomy focuses on an agent's desires and the motivational structure of their will. As I briefly touched on in the first chapter, Frankfurt's Humean approach to autonomy is often considered to be the more affective approach within the contemporary autonomy literature (Tappolet 2014, 165). However, as I also discussed in the first chapter, Frankfurt's theory of autonomy significantly differs from Hume with respect to the role emotions play in autonomy and the way in which we authorize and legitimize our autonomous choices. In fact, Frankfurt specifies that only "active love" can be an authoritative source of our actions, not any other emotions. While acknowledging that emotions have "motivational power," he states that they have no "motivational authority" (Frankfurt 1999b, 137). Rather, emotions can only move us by "sheer brute force" (Frankfurt 1999b, 137).

Importantly then, for Frankfurt, there is a distinction between being moved by our emotions and being constrained by what we care about—the difference being that while we can be "captivated by love," we are "enslaved by passion" (Frankfurt 1999b, 137). Similar to how he

conceives of compulsions being able to compel us to act, Frankfurt claims that emotions and impulses can be too powerful to resist but are independent of our own internal desires and intentions (Frankfurt 1999a, 136). Therefore, whereas active love helps us define our will by our essential nature as a person, constraining our actions with motivational authority; our passions are alien to us and lack such motivational authority. Consequently, for Frankfurt, the authority that active love has over our actions, as derived from one's own essential nature as a person, provides Frankfurt with a way to directly tie together our essential volitional characteristics with our actions. Thus, active love provides an answer to how we know our desires are truly our own (Frankfurt 1999a, 138).

However, by Frankfurt dismissing emotions as a source of autonomous action, claiming instead that emotions move us by "brute force," we are left with the idea that our emotions are somehow external to us and our own desires, acting as distractions that work against our autonomously willed decisions. Yet, as discussed throughout the second and third chapters, the empirical evidence does not support this idea. As summed up recently by Lerner et al. (2015), "…many psychological scientists now assume that emotions are, for better or worse, the dominant driver of most meaningful decisions in life" (Lerner et al. 2015, 801).

Indeed, we have known for quite a long time now that emotions are important when it comes to decision-making (Damasio 1994; Bechara et al. 1999; 2000; Wager et al. 2004; Litt et al. 2008). Numerous studies, as discussed in detail in Chapter Two, have shown that emotionally impaired individuals (e.g., those that have injuries to the vmPFC, like the well-known case of Phineas Gage) have difficulties in experiencing emotions, resulting in the inability to process risk and rewards adequately in decision-making (Damasio 1994; Bechara et al. 1999). Additionally, key network hubs like the vmPFC, that integrate cognitive, emotional, and motivational signals

have also been shown to be central to the decision-making process (e.g., in addition to the vmPFC, this also includes areas such as the ACC, insula, basal ganglia, and the amygdala, to just name a few). The decision-making research in the neuroscience literature, summarized throughout Chapter Two demonstrates that decision-making is not a cortical-based, highly cognitive process that is distinct from emotional and motivational processes. Instead, the research demonstrates that the decision-making process encompasses widely distributed areas of the brain (both cortical and subcortical), and involves cognition, emotions, and motivational processes like the reward network.

Many prominent accounts within the neuroscience of decision-making literature have pointed out the importance affect plays in decision-making, from biasing our initial information that we take in from within the decisional context, all the way to helping us compute the motivational value of our choices and directing our goal-directed behaviours. Less often discussed, Chapter Three also reviewed newer research on how emotion and motivation impact self-regulation, self-control, and cognitive control. As mentioned in the Frömer and Shenhav (2022) article discussed in Chapter Three, decision-making processes and cognitive control processes have important interactions that help support our behaviour, suggesting that goals affect the way we process stimuli (Frömer and Shenhav 2022, 3). Of importance to decision-making, they suggest that cognitive control helps direct attention to the attributes of our choice options that have the most heavily weighted value to us (Frömer and Shenhav 2022, 3). Frömer and Shenhav also point out that cognitive control plays an important role in directing the flow of reward information within the value network (Frömer and Shenhav 2022, 3).

Also discussed in Chapter Two and Chapter Three was research from Kable and Glimcher (2007), suggesting that regions in the brain that were initially thought to only be

involved in automatic reward responses, are actually involved in computing the value of both controlled and impulsive choices (Berkman et al. 2017, 424). Additionally, there is evidence that shows that brain activity from different areas of the brain also track the value of the various attributes of our choice options (e.g., gains and losses, emotional aspects, the value of waiting for a better outcome) (Berkman et al. 2017, 424). Insights from Clithero and Rangel (2014) also indicate that the subjective value of choice options are tracked in various value-related areas of the brain, such as the ventral striatum, ventromedial prefrontal, and orbitofrontal cortices (Berkman et al. 2017, 424).

Overall, as reviewed in detail throughout previous chapters, the research on decision-making suggests that emotional and motivational information is of crucial importance to the decision-making process, as well as our control processes that help support behaviours like making decisions and keeping our goal-directed behaviours on track. When we view these processes from a "whole brain" perspective, we can see the many different areas and networks interacting and integrating to ensure these processes are working.

However, this approach to decision-making and control, one in which emotional and motivational information is highly integrated into these processes, is rather different than Frankfurt's approach. As we saw, Frankfurt sees our emotions as detracting from our volitional choices, moving us by brute force instead of as necessary contributions of value-oriented information. While emotional information can certainly distract us when making decisions in certain contexts, it can also help us make better decisions in many other contexts. Importantly, emotions provide crucial information that we need in order to make a decision and are not separate and "alien" from the rest of our cognitive and motivational processes. Without the affective information, we would not be able to make good decisions for ourselves.

Furthermore, when we look at the kinds of issues commonly discussed as failures of autonomy (e.g., risky decision-making, impulsivity, disinhibition), they often stem from a problem within widely distributed networks that support our decision-making and control processes. For instance, as discussed in Chapter Three, research on the neural networks that mediate some of the decisional impulsivity behaviours, have suggested that interactions between the striatum, prefrontal cortex, and hippocampus are of central importance. Also important is the neuromodulation of the ascending monoamine systems (which includes dopamine (DA), serotonin (5-HT), noradrenaline (NA), and histamine (HA) circuitries (Dalley and Robbins 2017, 159). In particular, the nucleus accumbens (NAc), which as discussed in the previous chapter, receives dopamine (DA) input from the ventral tegmental area (VTA), is considered a central structure for impulsivity (Dalley and Robbins 2017, 159). Again, we can see that when these areas that are of central importance to value-based decision-making are compromised, we get problematically impulsive behaviours. Many of these areas are where cognitive, emotional, and motivational signals are integrated.

In terms of risky decision-making, there have been numerous studies linking damage to the vmPFC with this behaviour. As stated recently by Lerner et al (2015),

> Participants with vmPFC injuries repeatedly select a riskier financial option over a safer one, even to the point of bankruptcy in a game with real money, despite their cognitive understanding of the suboptimality of their choices. Physiological measures of galvanic skin response suggest that these participants behave this way because they do not experience the emotional signals—somatic markers—that lead normal decision makers to have a reasonable fear of high risks. (Lerner et al. 2015, 802)

As Lerner et al. (2015) point out, in addition to many others, problems in our emotional processes have a direct impact on the way we make decisions. They are also crucial for many processes related to self-regulation. This is an important distinction between the empirical

research on decision-making and self-regulation, and Frankfurt's theory of personal autonomy. Failures in emotional processes, or damage to value-oriented areas of the brain disrupt our normal mechanisms for decision-making and control. Therefore, Frankfurt's claim that emotions can only move us by brute force, rather than contributing to our autonomous decision is not compatible with the empirical research that suggests that emotions crucially help us decide what we do, in fact, *really* want to do. This distinction is quite significant and therefore presents a noteworthy tension between the two approaches.

6.2 Desires, Goals, and our Motivation System

Frankfurt's theory of personal autonomy is focused on the motivational structure of the individual. That is, of central importance to Frankfurt is a person's desires and motivational system—how do we decide which desires to act on? This is an important question for researchers examining the neuroscience of decision-making and goal-directed behaviours as well. However, there are some key differences between Frankfurt's approach and the neuroscience reviewed in previous chapters and the way central concepts like desires, goals, and motivation are being defined. I will take a closer look at these differences in this section. In particular, there are two main tensions between the two approaches that I will discuss in more detail. First, considering Frankfurt's earlier hierarchical approach, I argue that the neuroscientific conception of rewards and Frankfurt's lower-level desires function in importantly different ways. Second, in Frankfurt's more recent caring approach, I argue that there is an issue with his requirements for agential authority that needs some further exploration in light of the neuroscience of goal-directed behaviour and control.

6.2.1 Frankfurt's Hierarchy of Desires and the Reward System

Frankfurt's earlier conception of autonomy was centered around a motivational structure consisting of a hierarchy of desires, where first-order desires must be endorsed by second-order desires. The ability to hold higher-order desires is essential for personhood and is, therefore, how we derive agential authority over our actions. When comparing this with the neuroscientific account of our motivational structure and the way desires and goals motivate our actions, there are some key differences. However, to examine the differences between this account and the neuroscientific account, it would be useful to first define what is meant by the terms goals and desires, and the way in which desire and motivation operate in the neuroscience of goal-directed behaviour.

Generally, from an empirical perspective goals are defined as "mental representations of desired end-states" (Austin and Vancouver 1996, 338), with states being understood as outcomes or behaviours (Custers et al. 2012, 233). In other words, goals represent what we want (Custers, Eitam, Bargh 2012). The mental representations of goals help guide our behaviours toward these desired outcomes by controlling and motivating our actions toward them. While goals can vary in their level of abstractness, they all require goal-directed, instrumental actions in order to accomplish them (Custers, Eitam, Bargh 2012, 233). Goals can vary widely, from the simple goal of eating lunch to a complex, abstract future goal, such as a plan for your career. To further complicate this, in real life, people have many different goals operating at the same time: goals to get healthy, to further our career, to spend time with loved ones, etc. While some of these goals are compatible and can help support each other, others often conflict.

A relevant question for Frankfurt's approach, considering his focus on how we decide to act on one desire over another, is how these goals are formed and then selected. According to a

recent review on conscious and nonconscious goal-directed actions by Custers, Eitam, and Bargh (2012), the goal process starts with many options for goal pursuit that are either given (e.g., such as having to decide between driving or walking to get your coffee at a local coffeeshop) or generated (e.g., what am I going to do this afternoon?). Theories vary on how information is used in order to select the goal that is pursued, with some earlier approaches emphasizing conscious, deliberate selection (See for example, Deci and Ryan 1985; Gollwitzer 1990; Locke and Latham 2002), while some more recent theories argue that both conscious and nonconscious goal selection can often occur (See, Bargh 1990; 1997; 2006; Bargh and Ferguson 2000; Bargh and Gollwitzer 1994; Huang and Bargh 2014; Aarts and Dijksterhuis 2000; Aarts and Custers 2012). Despite these differences, most theories do share the idea that goals are selected based on their motivational value (Custers et al. 2012, 233). As Custers, Eitam, and Bargh further explain,

> The expected value principle…holds that the motivation to realize a certain outcome of state is the product of its value and the expectancy of being able to realize it. Hence, whether a particular outcome or state is set as a goal depends on its value or desirability and the expectancy of being able to attain the goal, or its feasibility. (Custers, Eitam, Bargh, 2012, 233)

The desirability or feasibility of the goals can be consciously deliberated, something that may occur particularly with novel goals (Custers, Eitam, Bargh 2012, 241). However, many of our goals are regularly pursued every day and are not novel. It has been argued that in those regular, everyday cases, both the desirability and feasibility, as well as the usual means to achieve the goal, can actually become integrated into the mental representation of the goal through our learning processes (Aarts and Dijksterhuis 2000; Bargh 1997; Bargh et al. 2001). As explained by Custers, Eitam, and Bargh (2012),

> As a result, conscious deliberation is no longer necessary and drops out of the equation, as priming a goal representation also activates information about

> desirability and will result, through spread of activation, in activation of the representations of the means we usually select to pursue the goal in a given situation. (Custers, Eitam, Bargh, 2012, 241)

Goal representations remain active in our working memory and can direct our behaviour and motivate our actions in flexible ways, suited for the constantly changing circumstances in our everyday life (Aarts and Custers 2012, 232).

In addition to the goal representation, or the actual setting of the goal content, equally important is the goal's regulation process (Moskowitz 2010, 140). The regulation aspect of goals is particularly important when it comes to implementing goals. However, it is important to note that in everyday life, people do not pursue one goal at a time. They often have multiple goals that need to be considered. Some of these goals may be compatible, but many times goals will also directly conflict and therefore, cannot be pursued at the same time (Cavallo and Fitzsimons 2012, 268).

Some goals may need to compete for valuable resources in order to be accomplished, while others may directly conflict in that the pursuit of one goal may end up impeding the progress of the other goal (Cavallo and Fitzsimons 2012, 268). This seems to be what is being described in Frankfurt's unwilling drug addict situation. On the one hand, the unwilling addict has the long-term goal of staying sober. On the other hand, they have the highly desirable short-term goal to take the drug. These two goals are motivationally, directly in conflict. This leaves the unwilling addict in the position where he needs to resolve this tension somehow between the two competing goals.

Traditionally, in the philosophical literature, temptations would be understood as desires, not goals. However, if we consider Frankfurt's classic unwilling drug addict example, where the unwilling addict has a conflict between two lower-level desires, I suggest that it seems less like a

desire conflict and more like a goal conflict, similar to those discussed in Chapter Three. From this perspective, it may seem odd to say that a person with a substance use disorder would have the "goal" of taking drugs, rather than a "desire" to take drugs. However, with the definition of goals being a representation of something you want, it would follow that wanting the drug is actually a goal for those with an addiction. Goals can be both adaptive, and in the case of people with addictions, they can also be maladaptive. We could say that, in the same way someone has a goal of exercising regularly (i.e., an adaptive, healthy goal), they could alternatively have the goal to take drugs regularly (i.e., a maladaptive goal, in that this goal is not going to support their well-being). Whether you end up prioritizing and pursuing certain goals, will then have to do with the desirability of the goals and how attainable they are—not necessarily whether they are adaptive or maladaptive.

The desirability of goals has to do with how rewarding or aversive they are to you. In other words, it is the motivational value of the goal that is the desire, not the goal itself. This definition of a desire is in line with the recent reward theory of desire proposed by Timothy Schroeder (2004), which suggests we can understand desires as representative of "rewards" (Schroeder 2004). This theory was also slightly expanded by Paul Thagard (2006), from which we can take a definition of desires as,

> To have an intrinsic (positive) desire for y is to use the capacity to perceptually or cognitively represent y to constitute y as a reward. Here y can range equally over states of affairs, objects, or features of objects, depending on the representational capabilities of whatever animal has the desire. (Thagard 2006, 153)

That is, from within the motivational system, we can understand rewards to be acting as desires. And the motivational value of these rewards impacts which goal we select. One of the main reasons that desire-based theories are considered more compatible with a naturalized

account of our decision-making processes is because of this acknowledgment of the motivational source of our autonomous actions as being affective- or desire-based. As discussed throughout Chapter Two and Three, affective-motivational information strongly contributes to the selection of our actions. This could lead one to conclude that affective, desire-based accounts could be supported from the standpoint of neuroscience research on decision-making. However, Frankfurt's approach, in distinguishing between higher-order and lower-order desires, ends up diminishing the role that lower-order desires (or rewards) play in guiding our actions. For him, it is the higher-order desires that are linked to our higher order volitions and therefore play the important action-guiding roles, as well as helping to legitimately authorize our autonomous actions (Frankfurt 1971).

Looking at Frankfurt's earlier hierarchical accounts of autonomy, he includes first-order desires as a source of motivation. Yet, Frankfurt also relegates first-order desires to being on par with temptations and impulses, all of which lack motivational authority (Frankfurt 1971). His early hierarchical theory of autonomy is focused on conflicts between these types of low-level desires, which lack any sort of authority in and of themselves. Instead, he determines that we need a higher-order desire to endorse one of the first-order desires. However, as we saw throughout the discussion of our motivation/reward system, we do not have different levels of rewards, some of which being more internal to our sense of self than others. There can be different levels of motivational value attributed to our desires, which ultimately impacts the choice we end up pursuing; but this still does not amount to what Frankfurt has in mind with higher-order desires.

Furthermore, the empirical research on the reward system shows that it is the first-order desires, or rewards, that do the bulk of the work motivating and guiding our actions through their

influence on our goal selection. More than just providing the motivational source in initiating actions, rewards also help guide and direct our attention and our actions, as well as keeping us on track with our goal-directed actions. Rewards also help us decide between goals when they are competing or conflicting, as well as help us prioritize certain goals over others. As discussed Aarts and Elliot (2012), "contemporary cognitive approaches to human behaviour propose that a substantial component of action is directed by goals that reliably control and motivate the behavioral system in a dynamic world" (Aarts and Custers 2012, 47).

In sum, when comparing Frankfurt's hierarchical theory of autonomy to the neuroscience research on motivation, we can see some important differences. Notably, the way both approaches understand concepts of goals and desires, and the way they operate in the motivational system are significantly different. First, I argued that the unwilling drug addict example should be seen as a conflict between goals that compete in terms of their desirability and feasibility, rather than a first-order desire conflict. Second, I argued that Frankfurt's structural account does not acknowledge all the action-guiding work that first-order desires/rewards perform in our everyday life, guiding and controlling many of our autonomous decisions and behaviours. Furthermore, the research on the reward system does not support the idea of a hierarchy of desires in the way that Frankfurt's early approach to autonomy would require.

6.2.2 Frankfurt's Caring Approach to Autonomy: Essential Volitional Character as the Source of Authority

As discussed in Chapter One, it has been argued that there was a significant shift from Frankfurt's early hierarchical account of autonomy to his later caring approach. This shift

represents a change from a conception of autonomy that is centred on *akratic* self-control to a theory of authenticity (Velleman 2006; Arpaly 2003). There is a key difference in the way Frankfurt understands the self-governing requirements from his early work to his later work and in particular, there is an important change in how he determines agential authority over actions. Frankfurt's earlier hierarchical approach to autonomy used higher-order desires as a way to endorse and legitimize an agent's authority over their actions, whereas his later caring approach focuses instead on an agent's essential character and the volitional unity of the will as an authoritative constraint on autonomous actions.

Frankfurt's early work required an agent to be governed by desires that were internal to the self and was heavily focused on dealing with examples of weakness of will (e.g., the unwilling drug addict example). In his most recent work, he changes his criteria for determining whether our desires are internal. Instead, he argues that motives are internal to the self when they are essential to an agent's volitional character. For Frankfurt, there is an important difference between active and passive love, where a person's interests are determined to be either conceptionally essential or inessential to one's will. If conceptually inessential, then they are considered separate and external to one's will (Frankfurt 1999b, 133). If a person is acting on considerations external to their will, then they are being "acted upon" (passive love), rather than their will determining their actions itself (active love).

This new conception of active love refers to a different type of self-government. Rather than control over one's desires, he requires a type of self-regulating motivational constraint—one that is tied to essential character. Frankfurt's caring approach introduces a type of "active self," a self that needs to have a hand in determining whether our actions stem from the essential character of our will in order to determine agential authority.

When comparing the empirical research on decision-making and how goals operate within our motivational system to Frankfurt's caring approach, there are some key differences that suggest Frankfurt's caring approach may not be supported by the empirical research as well. Frankfurt's introduction of this type of "active self" is not required in our decision-making and self-control processes. As discussed in the previous section, we can have goal conflicts without the chosen goal needing to say something special about the self. The desirability and feasibility of our goals are the main components driving our goal-directed behaviour. Therefore, it is the desirability and the feasibility of the goal that helps us decide whether to pursue the goal, not whether or not our desires are internal or external to our will. Our choices do not need to be based on whether a desire is somehow conceptually essential to our self, but rather we select from our choice options that have the most motivational value in that particular context.

While both Frankfurt and the empirical research on decision-making and control highlight the importance of "value" in our choices, I would argue that it is not the same type of value. Frankfurt links value to our essential personhood—that these deep values we hold reflect something special about the character of our will and therefore has the authority to constrain our actions legitimately. However, the neuroscience of decision-making relates value mostly to motivational value and the desirability of goals, biasing our choices. An object can have a high level of motivational value without it being meaningfully important to an individual (e.g., the unwilling addict being highly motivated to take drugs).

As previously discussed, goals will also come in and out of focus according to how desirable they are (their motivational value), their feasibility (are they more or less attainable), and their salience (if your attention is directed toward achieving that goal) (Cavallo and Fitzsimons 2012). Therefore, there does not need to be an active self—authentic or otherwise—

authorizing the selection between goals when they are in conflict. In fact, it has been persuasively argued that goals can compete on their own without conscious awareness (Huang and Bargh 2014), making the selection the motivational value or salience of the goals, rather than a governing, authentic "self" behind the choices.

There have been extensive studies looking at the various ways people overcome goal conflicts that go beyond a conscious active-self. Particularly, through a cognitive process like goal shielding, it has been suggested that goals can be maintained by inhibiting the activation of other goal possibilities, thereby ensuring the resources required to maintain the high-priority goal (Cavallo and Fitzsimons 2012, 273). This process can also become automated and nonconscious, as further discussed by Cavallo and Fitzsimons (2012),

> Because such goal conflicts are repeatedly encountered in a wide variety of self-regulatory pursuits, ignoring alternative goals, and maintaining focal goals are well practiced and become automated. Thus, the mere activation of certain goals may prevent the activation of alternative goals nonconsciously. (Cavallo and Fitzsimons 2012, 273)

How do we maintain our focal, or high-priority goals when they are in conflict with a highly desirable, more tempting goal (e.g., in self-control dilemmas like Frankfurt's unwilling addict example)? Research suggests that there seem to be strategies to delay gratification of more salient, short-term temptations in order to maintain conflicting long-term goals (Cavallo and Fitzsimons 2012, 275). It has been suggested that some people develop associative links between the conflicting temptation and long-term priority goal (Cavallo and Fitzsimons 2012, 275). The associations are suggested to develop in successful regulators from repeated activations of the priority goal when exposed to the temptation goal. For example, as discussed in Chapter Three, when dieters are tempted by unhealthy treats, they remind themselves of their long-term healthy eating goals (Fishbach, Friedman, and Kruglanski 2003; Kruglanski et al. 2002) Over time, this

association is strengthened and forms automatic cognitive associations between the temptation itself and the diet, automatically activating the higher-priority goal (Cavallo and Fitzsimons 2012, 276). Interestingly, these studies suggest that "temptations" can actually help us reinforce our long-term goals rather than acting as detractors, as suggested by Frankfurt's unwilling addict example.

Also, as discussed in Chapter Three, newer definitions of self-control include various types of goal conflicts that go beyond the short-term reward versus long-term goal conflict. The expanded definitions also include other ways to resolve these goal conflicts that go beyond the effortful, conscious inhibition of temptations. The reframing of self-control in this way also moves us further away from the traditional idea that self-control necessarily involves a conscious, "active self," operating as the enforcer of our self-control processes (Inzlicht et al. 2021). According to the expanded approaches to self-control, a conscious, active self is not always needed. There are other ways in which we can resolve goal conflicts that go beyond conscious inhibition, strategies that do not involve conscious effort. These strategies may also be more proactive, rather than inhibiting temptations after they arise.

Therefore, there are important differences between the neuroscience of decision-making and control and the way goal-directed behaviours are understood, and how Frankfurt conceives of desires and motivation in his caring approach to autonomy. One key difference is the fact that goals can and should compete and operate on their own, without being controlled by an essential self. In fact, to function effectively, we need to be able to evaluate things like the motivational value of our choices and goals quickly and often nonconsciously.

Overall, when we consider desires and goals and how they operate within the motivation system, there are some important differences between the way Frankfurt conceives of these

concepts, and the way that current neuroscience explains the function of our motivation system. Primarily, there are tensions between Frankfurt's idea of higher-order desires providing the authority over our choices, when in fact rewards, or lower-order desires do much of the work in guiding our actions. Additionally, the empirical literature on self-control has been moving away from definitions of self-control that involve a type of "active self" as the source of control over our actions, one that has to consciously and effortfully resist temptations and distractions from our long-term goals. At the heart of these differences between the neuroscience literature and Frankfurt's two approaches to autonomy lies a key difference in the way that the value of an object is understood. Frankfurt's conception of value is tied to an agent's essential volitional character, which directs and constrains our autonomous actions. In contrast to this conception of value, a neuroscientific approach to our motivational system would suggest that the motivational value of our choices primarily controls and motivates our behavioural choices.

6.3. The Unwilling Drug Addict and the Neuroscience of Addiction

Frankfurt uses the unwilling addict example, or variations of this case, throughout his work as a paradigmatic example of an autonomy conflict. For Frankfurt, the unwilling drug addict's autonomy is compromised because he is acting on a desire that is external to him. Or, in other words, the addiction is guiding his decision rather than the desire which he has willed to be effective. However, for Frankfurt, the unwilling addict is still a person capable of higher-order desires and volitions. While the unwilling addict may make a heteronomous choice to take the drug, he is not necessarily altogether lacking the capacity for autonomy.

However, the empirical research on addiction suggests that a person with a serious addiction would have their autonomy impacted in many different ways, not just in the specific

and very localized way Frankfurt uses this example to demonstrate an autonomy conflict. It is well documented within the neuroscientific literature that there are certain neuroadaptations that occur in the brain when someone has an addiction (Volkow et al. 2019), changes that can affect one's capacity for self-regulation and decision-making (Koob and Volkow 2010). We know from the neuroscience research on addiction that it is a chronic, relapsing disease, which can occur when those who are vulnerable (due to genetics and developmental and adverse social exposures) are repeatedly exposed to the drug. However, as Volkow et al. (2019), summarizes below,

> As of now, it is not clear what determines the transition from drug experimentation to addiction, which emerges when individuals lose their ability to overcome the strong urge to take the drug despite a conscious awareness of not wanting to do so and the recognition of their potentially catastrophic consequences. However, we do know this transition is associated with measurable disruptions in several brain circuits including those involved with conditioning, reward sensitivity, incentive motivation, self-monitoring/regulation, mood, and interoception. (Volkow et al. 2019, 2121)

Volkow further explains that addiction impacts the reward circuit's capacity to appropriately respond to rewards and thereby motivate actions that are not drug related, such as other unrelated goals (e.g., career goals, or interpersonal goals). Addiction also enhances the emotional stress circuits and impairs the capacity to self-regulate. These changes result in compulsive drug-seeking and drug-taking, as well as an inability to control one's cravings to take the drug—even with a strong desire to quit (Volkow et al. 2019, 2116). Addiction can also result in the kind of behaviours often seen with severely addicted individuals, behaviours such as over-prioritizing seeking and taking the drug at the expense of any other important concerns (e.g., being incarcerated, losing custody of children, losing your career, etc.). Or, in other words, the changes result in the goal-directed system in these individuals getting skewed, where the

motivational value of their goals changes in ways that are problematic for their general everyday functioning and well-being.

We can understand some of these issues as a type of reward/value system dysfunction, where the circuits acting as a regulatory mechanism underlying the reward network—the prefrontal brain regions that normally regulate reward values by modulating dopamine increases in the ventral striatum—can stop functionally properly in people with and addiction (Koob and Volkow 2010, 230). As Volkow further explains,

> …human neuroimaging studies show that the prefrontal cortex (orbitofrontal, medial prefrontal, prelimbic/cingulate) and the basolateral amygdala are critical in drug-and cue-induced craving in humans. In prefrontal regions (e.g., cingulate gyrus and orbitofrontal cortex), these changes have been associated with a reduction in striatal dopamine D2 receptor availability observed in addicted subjects. These associations could either reflect a disruption of frontal brain regions secondary to changes in striatal dopamine activity, or alternatively they could reflect a primary disruption of frontal regions that regulate dopamine cell activity. Indeed, a recent PET study provided evidence that prefrontal brain regions regulate the value of rewards by modulating dopamine increases in the ventral striatum, a regulatory mechanism that becomes dysfunctional in addicted individuals. (Koob and Volkow 2010, 230)

The types of neuroadaptations described require a reassessment of the capacity for autonomy in a person with addiction. The unwilling addict, or even the willing addict, would both have their capacity for autonomy impacted much more widely than Frankfurt's initial analysis. What undermines their personal autonomy is not just a matter of acting on a desire that they do not endorse as internal to themselves in the way that Frankfurt outlines. Rather, the fact that addiction also causes changes in cognitive, emotional, and motivational processes which, in turn, can "cause the over-valuing of drug reinforcers at the expense of the under-valuing of natural reinforcers, and deficits in inhibitory control of drug responses" (Koob and Volkow 2010, 230). That is, the neuroplastic changes that occur with addiction more generally affect the

way they make decisions, the way their reward system assigns value to goals (over-prioritizing drug-related goals and under-valuing goals not related to drug-seeking or drug-taking), as well as impacting their ability for self-regulation and inhibitory control. These are crucial elements for the capacity for personal autonomy, as was shown in Chapter Two and Chapter Three. Disruptions in decision making, self-regulation, and inhibitory control processes can have a profound effect on one's ability to make autonomous decisions, or to be able to act in the way they want to act.

Therefore, the example of the unwilling addict and the question of their capacity for personal autonomy is quite complex. Following from the neuroscientific research on addiction, there is an autonomy issue present in a person with an addiction; however, it is much more pervasive than Frankfurt's discussion of the unwilling addict. Much more than just causing an action that is in some sense external to them, an addiction can cause many changes to a person with addiction's actual decision-making, self-regulatory, and inhibitory control processes which would inevitably lead to many different types of autonomy issues (e.g., issues of impulsivity, problems in decision-making, etc.). Therefore, a person with an addiction would likely have their autonomy compromised no matter the desire they choose to act on, whether they give in to the craving or not.

6.4 The Subjective Experience of Agency and Control

As discussed throughout earlier sections in this chapter, we may not always be consciously aware of our specific motivating desires in much of our goal-directed behaviours. However, there is some interesting research looking at certain conscious experiences related to our intentions to act and control our actions, as reviewed in Chapter Four. The research on our

subjective feelings of agency and control may have specific implications for Frankfurt's requirement for agential authority over our actions. As previously discussed, Frankfurt requires that an agent "feel satisfied" with their desires. However, how reliable are these kinds of subjective feelings about our desires and choices? This section will consider this question and whether there are any tensions between Frankfurt's requirements for autonomy and what we've learned from the neuroscience research underlying our feelings of agency and control over our actions.

Recent studies on the sense of agency and motivation literature have provided more information about the different subjective feelings associated with our intentional action choices (Desmurget and Sirigu 2012; Berridge et al. 2009; Haggard 2008). These include feelings of "wanting to move," feelings of "motor urges," unconscious feelings of "wanting" and "liking," and the sense of agency, ownership, and control over our actions—all of which seem to help inform our experiences of choosing to do what we want to do. For instance, Desmurget and Sirigu (2012) have proposed that our goal directed actions can trigger two different types of subjective experiences of conscious intentions, mediated by two different areas of the brain. First, the inferior parietal lobe (IPL) involves the experience of 'wanting to move', which provides a general goal of action before planning a specific movement (Desmurget and Sirigu 2012, 1004). The second kind of subjective experience comes from the mesial precentral area (MPL), which includes the supplementary motor area (SMA) and the pre-supplementary area (pre-SMA). This area mediates a feeling of 'motor urge,' which signals the time when the planned motor action is about to begin (Desmurget and Sirigu 2012, 1004).

According to Desmurget and Sirigu, these two areas mediate both experiences involved in giving us the feeling of conscious intentions in our actions. These areas are part of wider

networks related to the generation of our voluntary actions, mediating our motor planning, and action control more generally. The wider network areas include the posterior parietal cortex, the anterior cingulate cortex, the supplementary motor area, the dorsolateral prefrontal cortex, and the basal ganglia (Desmurget and Sirigu 2012, 1005).

In addition to feelings of conscious intentions, we can also experience feelings of agency, or the experience that one's own actions caused a specific outcome in the external world (Haggard 2008, 942). Haggard outlines the distinction between feelings of conscious intention and agency as, "agency necessarily involves the experience of external sensory consequences, whereas intention relates more closely to preparation and effort" (Haggard 2008, 942). Haggard and colleagues, after doing years of experiments in this area, still do not have a definitive answer to what underlies the specific phenomenon of the "feeling of agency." What their studies do seem to suggest overall is that "explicit attributions of sensory effects to one's own action or to another agent involve the parietal cortex, but the normal, implicit sense of one's own agency might involve predictions derived from intentions that are generated in the frontal motor areas" (Haggard 2008, 942).

Additionally, the Fried et al. study (1991), differentiates between the conscious feelings associated with the SMA versus the parietal area of the brain. This study showed that intentions triggered by stimulation in the supplementary motor area (SMA) produced the feeling of a compulsion, in contrast to those triggered by stimulation of the parietal area which produced the feeling of voluntary intentions to act (Fried et al. 1991). Furthermore, interesting results from the study on the relationship between coercion and a sense of agency by Caspar et al. (2016) showed that obeying instructions reduces the sense of agency (Caspar et al. 2016), "producing an experience that is closer to passive movement than to voluntary action" (Haggard 2017, 206).

These results are consistent with the idea that we feel less control over actions that we are coerced into doing, rather than those that are freely chosen without coercive influence.

An important area known to play a key role in the experience of our conscious feelings of cravings is the insula. The insula (especially the more anterior regions) has reciprocal connections to several cortical and subcortical regions (such as the ventromedial prefrontal cortex, the amygdala, and the ventral striatum, parts of the motivational system). The insula seems to play a key role in interoceptive functions (i.e., relating to stimuli from within the body), where it helps integrate autonomic and visceral information with emotion and motivation (Koob and Volkow 2010, 232). The insula seems to be crucial in providing our conscious awareness of urges or cravings, as well as providing more general awareness and understanding of the stimuli coming from internal stimuli within the body (Naqvi and Bechara 2009).

6.4.1 Experiencing "Autonomy Disturbances"

When there is an issue within these networks, problems with how we perceive our own actions can occur. For instance, in cases of alien hand syndrome, or other similar utilization behaviour disorders, issues related to the parietal-premotor circuit seem to contribute to this type of "autonomy disturbance." The parietal lobe and the pre-SMA seem to be key areas contributing to the experiences that these patients have when acting. The parietal-premotor circuit is also a key area for our subjective feelings of voluntary action and control. It seems likely that these areas play crucial roles in both keeping our voluntary actions focused on the task at hand as well as contributing to our feelings of agency in our actions and control over our bodies (Haggard 2008, 939).

For example, patients with utilization behaviours (due to bilateral lesions on the SMA) will be unconsciously and automatically triggered by stimuli in their immediate environment, resulting in their hand involuntarily and inappropriately grasping objects around them without any conscious awareness or intention (Hassan and Josephs 2016). This is often referred to as a disorder of personal autonomy for obvious reasons. These individuals do not want their hand to be acting in the way that it has acted, nor do they have any reasons to have grasped the particular object. When asked about it after, they are often surprised about the fact that they grasped the object to begin with. In cases like these, we can see how a disturbance in the circuits related to the experience of our voluntary actions can contribute to feeling like your action was not your own (Hassan and Josephs 2016).

In cases where an individual is acting in a goal-directed way, external stimuli in their environment can trigger unconscious wanting of a reward, thereby triggering the goal-directed behaviour. For example, unconsciously perceiving an apple can trigger one's goal to eat a healthy lunch, leading to the completion of this goal. So why would this type of unconscious environmental influence on behaviour not be considered autonomy undermining in the same way that it is in the cases of utilization behaviours?

The crucial difference in these two cases is a person's capacity to prevent all non-relevant environmental stimuli from affecting one's action choices. A patient with utilization behaviour disorder lacks this ability to inhibit the irrelevant, or non-task specific stimuli that they encounter in their environment—leading to improper behaviours that they did not choose to perform. Whereas those who do not have this issue with their inhibitory control mechanisms would unconsciously be able to refrain from reacting to every aspect in their environment. If the apple was not in fact relevant to any of their goals, then they would not have reacted to it. Someone

with a utilization behaviour syndrome would not be able to refrain from grasping the apple regardless of its relevance to their goals.

A paradigmatic case of lacking this capacity to block out irrelevant external stimuli are patients with the rare Environmental Dependency Disorder ("Zelig-like syndrome"), such as the example of the Chameleon Man discussed in Chapter Four. This particular type of loss of control severely impacts one's sense of identity and how the self interacts with the environment around it (Conchiglia et al. 2007). Without this type of control, it would be very difficult for someone with this disorder to maintain their personal autonomy consistently, over time. A "chameleon-like" person would constantly be shifting their actions based on the environmental stimuli around them, rather than holding consistent self-directed goals across longer periods of time. In this sense, it is a type of personal control that keeps a person able to separate their own actions from undue environmental influence that would lead to actions that are not wanted. And this type of self-regulatory control is not necessarily something in which we would be explicitly consciously aware.

6.4.2 The Experience of Agency and Control: Implications for Frankfurt

The intuition that our behaviours originate in our own conscious and intentional choices is not only prominent in many accounts of personal autonomy but also in many traditional psychological approaches to understanding motivation and action. These traditional theories assume a conscious agent is in control—making effortful, deliberate judgments, deciding on courses of action, and ensuring the follow-through on those intentions (Deci and Ryan 1985; Baumeister, Tice, and Vohs 2018; Locke and Latham 2002; Mischel 1973).

Typically, as discussed in Chapter Five, the requirement of a conscious agent is more prevalent in reasons-responsive accounts of autonomy, such as Korsgaard's approach to autonomy. Frankfurt's account does not require overt conscious awareness of one's desires (Frankfurt 1971, 8); however, his approach does require a capacity to determine whether we are satisfied with our desires—a capacity that would be difficult to accomplish if we are not always consciously aware of our desires or unable to reliably know how we feel about our lower-order desires. As Frankfurt states,

> Satisfaction with one's self requires, then, no adoption of any cognitive, attitudinal, affective, or intentional stance. It does not require the performance of a particular act; and it also does not require any deliberate abstention. Satisfaction is a state of the entire psychic system—a state constituted just by the absence of any tendency or inclination to alter its condition. (Frankfurt 1999c, 104)

Frankfurt's approach requires this ability to determine how we feel about our desires and to recognize this state of satisfaction signifying our autonomy. Do our desires feel alien to us? Do we endorse these desires and feel satisfied with internalizing the desires as our own? Or do we still feel unsatisfied with our choice? What exactly do our motivational desires and our experiences of acting autonomously feel like to us?

I will consider two main tensions between Frankfurt's account and the empirical literature on these types of subjective feelings related to our desires. First, if we are not always consciously aware of our desires before choosing to act on them, does it matter for autonomy that we feel a certain way about them? And second, can we then, consequently, rely on this information in legitimizing our agential authority over our actions?

We do not have all the answers yet in terms of understanding exactly how consciously aware we are of the processes behind many of our everyday actions and choices. However, as we

saw throughout the research in previous chapters, there is a lot of empirical support suggesting that conscious awareness is not always at the forefront of much of our day-to-day goal-directed behaviours. The research reviewed has shown that much of our motivational system underlying our goal-directed behaviour is often, in fact, regulated and controlled by generally nonconscious processes (Bargh 1990; 1997; 2006; Bargh and Ferguson 2000; Bargh and Gollwitzer 1994; Huang and Bargh 2014; Aarts and Dijksterhuis 2000; Aarts and Custers 2012).

Additionally, the research covered in Chapter Four on voluntary action and our subjective experiences of intentionality and agency, also points to the possible conclusion that our actions likely do not originate in a conscious intentional decision to perform that action. That is, specifically, the experience of conscious intention seems to not be the cause, but rather a consequence of the motor preparation processes occurring as we get ready to act—meaning, "we do not act because we 'want to act' but we 'want to act' because the brain has been preparing for action" (Desmurget and Sirigu 2012, 1005). This suggests that our intentions to act do not always need to be conscious to lead to actions we would consider to be autonomous. Nonconscious motivational information can, and often does, shape many of our voluntary behaviours.

This presents a possible issue for Frankfurt in both his hierarchical and caring approaches to autonomy. Both approaches require a capacity to differentiate between desires that feel right and desires that feel alien to one's will. From the goal-directed perspective, we may not always be consciously aware of our motivations, but should that mean that the actions cannot be autonomous? We do not always need to be consciously aware of our intentions in order to accomplish highly effective goal-directed actions.

In terms of the reliability of the feelings we do have about our actions, they may not always be able to accurately tell us something about the autonomy of our actions either. For example, we can consider the previously discussed "Chameleon Man," whose entire personality, desires, and goals change according to the environmental setting in which he finds himself. The Chameleon Man, if asked, would likely say that he feels that the desires motivating him at any given time are desires that he feels satisfied with. However, we would not want to agree that he is acting autonomously given his paradigmatically non-autonomous actions.

This problem is also apparent in Frankfurt's discussion of the unwilling drug addict, where Frankfurt points out that the addict does not want to act on the desire to take the drugs and, in fact, he dislikes his addiction. For Frankfurt, the unwilling addict needs to determine which of the conflicting desires surrounding this decision is the desire that he really wants, and this is determined by how he feels about his desires. Does he *want* to want them? Does he like them? These are the key questions for the unwilling addict.

From Frankfurt's perspective, the fact that the unwilling addict does not like his desire to do drugs tells us something important about whether his decision to be moved by that desire is autonomous. However, studies on processes related to feelings of "wanting" and "liking" and the dissociation between the two processes (Robinson and Berridge 1993), may raise some questions about this. Robinson and Berridge (1993) have argued that there is a distinction between "wanting" (our unconscious motivational incentives), and our hedonic reactions of "liking" (the pleasure we experience from consuming rewards). Additionally, our pleasure reactions to rewards can be further differentiated as conscious liking, or unconscious "liking". The "liking" process, rather than being mediated by dopaminergic systems like our unconscious "wanting," uses discrete brain regions termed "hedonic hotspots." The hedonic hotspots are found in the

nucleus accumbens, ventral pallidum, orbital frontal cortex, and the insular cortices (Castro and Berridge 2014), which act on opioid and endocannabinoid neurotransmitters (Anselme and Robinson 2016).

While there can be overlap in the "wanting" and "liking" processes, research has suggested that these processes, both of which use the mesolimbic structures, can also be dissociated. This means that we can sometimes "want" what we "like," where we pursue rewards that we find pleasurable, or "want" what we do not "like," where we pursue rewards that we do not find pleasurable (Anselme and Robinson 2016; Robinson, Anselme, et al. 2015).

The motivation/reward network has been shown to be crucial in early decisions of whether to act or not and mediates our unconscious "wanting" (Haggard 2008, 938; Desmurget and Sirigu 2012). The motivational part of our motor circuitry relies on the basal ganglia network, or more widely on the mesocorticolimbic dopaminergic circuit, which mediates our experiences of unconscious "wanting" (Desmurget and Sirigu 2012, 1005). This type of "wanting" would be based on the processing of inputs from: (1) the internal state of the person; and (2) the reward values of their external cues (Desmurget and Sirigu 2012, 1005).

The dissociation between "wanting" and "liking," has been demonstrated reliably across many studies of addiction. People with substance use disorders, particularly with highly addictive drugs like opioids, show intense cravings for drugs (the drugs themselves causing unnaturally high levels of dopamine, triggering intensely strong cravings) while reporting either no increase in pleasure when the drugs are attained or sometimes even a decrease in pleasure experienced when consuming the drug (Anselme and Robinson 2016, 125). In these cases, the people with addictions strongly want the drug, but do not like doing the drug.

The dissociation of these two systems behind "wanting" and "liking" demonstrates how a person can both excessively want a drug and not like the drug at the same time, without making the conflict an issue of authenticity. However, for Frankfurt, the example of the desire conflict in the unwilling addict example requires us to rely on these types of subjective feelings about our desires to act, where our feelings about our desires is supposed to give us important decisional information. We need to be able to assess whether we feel alienated from a desire or feel satisfied with a desire. However, it may be that sometimes we just do not like what we want or want what we like. It could become difficult to try and detangle these different subjective feelings that can occur and certainly makes it difficult to rely on them in terms of providing key information about the authority of our choices.

To summarize, the empirical research on our subjective experiences related to agency and control highlights some potential tensions between it and Frankfurt's requirements for agential authority. First, there is evidence to suggest that we have often selected our choices nonconsciously before any conscious feelings arise in relation to this choice. Therefore, Frankfurt's requirement of "feeling satisfied" with our choices may not have as much of an impact on directing and providing authority over much of our goal-directed actions. Second, Frankfurt asks the question of how we know if a desire is one we really want to have? However, as suggested by the studies on wanting and liking, the reliability of such feelings being the determining factor in how we feel about our desires is questionable. The dissociation between our "wanting" and "liking" seems to impact this issue directly, particularly in the case of addictions. Therefore, feeling satisfied with our desires as a requirement for determining whether our choices are autonomous could lead to some problems.

Conclusion

Approaching the concept of personal autonomy from a neuroscientific perspective and considering the processes that underlie our decision-making, self-control, and feelings of agency highlights some interesting differences between the research and traditional theories of personal autonomy like Frankfurt's account. First, I considered the key differences between the integrated, cognitive-emotional-motivational frameworks for decision-making and self-control versus Frankfurt's approach to autonomy. There are significant amounts of empirical research demonstrating that when network damage affects the way emotional and motivational information is processed in decision-making (e.g., such as damage to the reward network), this results in objectively worse decision-making outcomes. Key hubs that help integrate cognitive, emotional, and motivational information in these networks are also crucial for self-regulation and self-control and therefore have a direct impact on our goal-directed behaviour. This research is fundamentally different from Frankfurt's approach, which is fairly dismissive of the role that emotions play in determining autonomous decision-making. Rather than acting as distractions from our autonomous decisions, or our will, emotions should be considered integral to processes involved in autonomy like decision-making and self-control.

Second, I looked at the way in which central concepts like desires, goals, and motivations are defined in both Frankfurt's theory and from a neuroscientific perspective. I suggested that there are some important differences between these two approaches. I first argued that there are some possible tensions between Frankfurt's hierarchy of desires and the way rewards operate within our motivation system. Frankfurt's requirement for agential authority over our choices is placed within the realm of higher-order desires, whereas I argue that the neuroscience suggests the bulk of the action-guiding work occurs in what Frankfurt would call the lower-order desires

or rewards. I also examined how the neuroscience of rewards and motivation impacts his later caring approach to autonomy. The empirical literature on self-control has been moving away from definitions of self-control that involve a type of "active self" as the source of control over our actions, one that has to effortfully resist temptations and distractions from our long-term goals. I argue that Frankfurt's caring approach ascribes to this sort of essential self as providing the motivational authority over our actions, which is not compatible with the neuroscientific research on the way we deal with goal conflicts.

The third section looked more closely at Frankfurt's unwilling drug addict example, comparing this with the recent neuroscience of addiction research. I considered Frankfurt's claim that the unwilling addict could have the capacity for personhood. Instead, I argued that a person with a severe substance abuse problem would likely have a reduced capacity for personal autonomy. The research on addiction demonstrates that the addiction can cause many pervasive changes to a person's decision-making, self-regulatory, and inhibitory control processes that would likely lead to many autonomy issues like risky-decision-making, issues of impulsivity, and generally a reduced ability to inhibit their drug-related activity. In particular, it was shown that addiction negatively impacts the functioning of the reward network, therefore damaging our rewards-based valuation system that I argue underlies our autonomy processes.

Lastly, I examined some potential issues related to how we experience feelings of agency and control versus Frankfurt's conception of the way we experience our motivations and act on our desires. I argued that there were two main tensions between these perspectives. First, there is strong neuroscientific support suggesting that we are not always consciously aware of our desires before choosing to act on them—with our goals often competing without our conscious awareness. Therefore, I question whether Frankfurt's account, with its requirement for feeling

satisfied with our desires, is plausible as a requirement for autonomy. Second, I examine Frankfurt's requirement of relying on these types of feelings for agential authority. In particular, I discuss the neuroscientific research on the motivation system, with possible dissociations between our feelings of "wanting" and "liking." I argue that these feelings of "wanting" versus "liking" could present a problem for Frankfurt's caring approach to autonomy and his requirement for agential authority.

Therefore, despite the sense that affective, desire-based accounts of autonomy could be more compatible with a naturalized account of autonomy, there are still some significant issues with Frankfurt's theory in light of the neuroscience research—issues with the role of emotions in decision-making, how we understand desires to motivate our actions, the role that lower-order desires play in guiding and motivating our actions, and with relying on our feelings of "wanting" and "liking" to help determine the authority of those desires over our actions. My hope with this chapter overall was to begin to point out some discrepancies between the recent neuroscientific research on decision-making, control, and our motivation system, and the way in which Frankfurt uses these concepts within his theory of personal autonomy.

Conclusion

The central aim of this dissertation has been to explore how recent neuroscience research might bear on prominent theories of personal autonomy, with a focus on the work of Korsgaard and Frankfurt. I examined the neuroscience of decision-making, self-control, voluntary action control, and the conscious feelings related to agency and ownership in order to widely capture the many processes that underlie personal autonomy. I argued that we need to rethink certain concepts that are central to personal autonomy, such as emotion, motivation, and desire and the ways in which they affect our understanding of decision-making and self-control. By scientifically updating these concepts, this in turn can update the way we think of these concepts from within the philosophical literature.

I examined some of the emerging trends in cognitive neuroscience that suggests that complex behaviours emerge from large-scale networks across widely distributed areas of the brain, where cognitive, emotional, and motivational information are deeply integrated. Importantly, cognitive, emotional, and motivational information have been shown to be highly integrated in processes of decision-making and self-control. My hope was to point out that within the neuroscience research, there was a rather large shift from viewing reason and emotion as hierarchical, separate, competing processes, to an integrated cognitive, emotional, motivational approach in decision-making and control. This shift impacts that the way we can think about autonomy as well. Additionally, I argued throughout that the neuroscience literature on decision-making and self-control provides plenty of empirical evidence supporting the conclusion that emotion plays a larger role in autonomy than is usually discussed within the philosophical literature. I argued that an affective, rewards-based valuation system seems to underlie many of

the autonomy-related processes discussed, where integrated cognitive, emotional, and motivational information help support key processes such as decision-making and control.

As previously discussed, several recent theories of autonomy have started to include emotion into their approaches, thereby moving these theories closer to being in step with current neuroscience. However, there are still some empirically outdated ideas circulating within the personal autonomy literature. Examples of these more outdated ideas that I address are: (1) the Platonic idea that reason, in an important sense, needs to be in control of our emotions; (2) the idea of an emotional decision competing with a rational decision; and (3) the idea that an agent needs to be consciously aware of their desires as a necessary requirement for autonomous action. These ideas, arising within Korsgaard's and Frankfurt's approaches, highlights some important tensions between the empirical literature and philosophical autonomy literature.

Chapter One provided a more historical basis for understanding how these ideas made their way into the autonomy literature. I argued that the idea that reason and emotion are somehow at odds when it comes to making decisions has a long and deeply entrenched philosophical history, established by the work of Plato, Kant, and Hume. I discussed contemporary theories of autonomy that are rooted in these historical approaches: the theories of Harry Frankfurt, Gary Watson, and Christine Korsgaard. I argued that all three contemporary approaches maintain, to a certain extent, the idea that emotions act as distractions from rational decisions and are generally dismissive of emotions playing a key role in autonomous decision-making.

Chapters two through four turned to the neuroscience research, examining key information about how processes relevant to autonomy operate in the brain. Chapter Two examined research on the neural correlates of decision-making. In this chapter, I argued against

dual-process models of decision-making that pit cold, deliberative decisions and hot, emotional decisions against one another. I summarized the extensive amount of research that supports an integrated cognitive, emotional, motivational account of decision-making. These frameworks highlight the importance of affective processes like emotions and motivational rewards in making decisions and point out key hubs in the brain where cognitive, emotional, and motivational information is integrated within neural networks. This chapter demonstrates the affective basis of our valuational processes that play a key role in decision-making.

Chapter Three reviewed the neuroscience of self-regulation, self-control, and cognitive control processes. These types of control processes help support our decision-making and goal-directed behaviours. Historically, both in the philosophical and empirical literature, self-control was usually thought of as a highly rational process—a process where we consciously, and effortfully, override our emotions and impulses to achieve our goals. However, Chapter Three argued that emotional and motivational information is a crucial component of control processes. I reviewed many accounts of self-regulation, self-control, and cognitive control that argue that emotional and motivational information is required, and in fact, that control may be a more affective process than we imagined.

Chapter Four looked at an interesting subsection of research focused on the neuroscience of volition. In this chapter I examined some key aspects of the research on self-generated voluntary action and our conscious feelings of agency and ownership over our actions. I suggested that, although this research is focused on the neuroscience of free will, it also had some relevant information for understanding the way the brain produces behaviours related to autonomy. In this chapter, I argued that the focus of Libet's and Haggard's experiments on trying to capture endogenous action by excluding external signals may be too artificial to capture the

way we make autonomous decisions in everyday life. Instead, I suggested that it may be important to look at how we make decisions within environments that have external environmental factors to really understand the way in which we manage external stimuli while making autonomous decisions. Importantly, this chapter also demonstrated the key role of the basal ganglia in supporting our voluntary action choices, where cognitive, emotional, and motivational information is integrated to effectively select between action options. Much of this motivational value is determined automatically and nonconsciously, which in I argue in later chapters is quite different from the traditional philosophical understanding of autonomous decisions as conscious, deliberative choices.

This chapter also looked at the related research on the conscious feelings of agency and ownership over our actions. This research suggests that the connectivity between frontal and prefrontal motor areas responsible for initiating voluntary action and the parietal areas that monitor our perceptual events may contribute to these conscious feelings of agency and ownership. This research relates to the way we experience our autonomy and understand our identities as being held constant over time, amongst various environmental signals.

Finally, in Chapters Five and Six, I considered the possible implications that the neuroscience of decision-making, self-control, voluntary action, and conscious feelings of agency may have on the theories of Korsgaard and Frankfurt. First, in Chapter Five, I examined the way emotion is discussed in Korsgaard's theory of autonomy. I argued that her conception of emotion and its role in decision-making is at odds with the integrated neuroscience accounts of decision-making. In particular, I claimed that "rational" decisions competing with "emotional" decisions, is not reflected in the neuroscience of decision-making. Rather, I argued that all decisions importantly include affective information.

I also pointed out that Korsgaard's account requires a conception of personhood that is tied to the Platonic idea of self-control—one in which an agent's reason must be in control of their emotions, motivation, and any impulsive temptations. I argued that this is a scientifically outdated conception of self-control. Instead, the neuroscience of self-control suggests that emotional and motivational information play a crucial role in control processes.

In the final section, I examined Korsgaard concept of "reflective distance," or the conscious deliberative process of reflecting on our desires. I argued that there is an important difference between Korsgaard's concept of reflective distance and how our motivation processes impact our decision-making. I argued that nonconscious goal-directed behaviours play a much larger role in our everyday life, and that these should not be considered heteronomous because they are a normal part of our processes. Overall, I concluded that her requirement of "conscious reflection" for agency is not compatible with neuroscientific conceptions of emotion and motivation, and how they contribute to decision-making and self-control.

Lastly, in Chapter Six I argued that there are also some important tensions between the neuroscience of decision-making, control, and our conscious feelings of agency, and Frankfurt's approach to autonomy. First, I argued that there are some fundamental incompatibilities between Frankfurt's approach to autonomy and the integrated, cognitive-emotion-motivational frameworks for decision-making. Decades of research supports the idea that problems in our emotional and motivational processing or damage to specific hubs that integrate cognitive, emotional, and motivational information in the brain have a profound impact on our decision-making capacities. These areas related to emotional and motivational processing have also been found to be crucial for self-regulation processes, which are necessary for goal-directed behaviour. This research differs significantly from Frankfurt's theory, which I argued, dismisses

the role of emotions in autonomous decision-making. Rather than acting as detractors of our willed actions, I claimed that emotions should be considered central to autonomous decision-making.

I also argued that there are crucial differences between the way central concepts like desires, goals, and motivation play a role in autonomy in Frankfurt's approach and the way they should be understood from a neuroscientific perspective. Frankfurt's conception of agential authority requires that higher-order volitions must guide our autonomous actions. However, I argued that the neuroscience of rewards and motivation suggest that lower-level rewards actually do the bulk of the work in guiding our goal-directed actions. I also pointed out that Frankfurt's conception of an essential, "active self" as the source of control over our actions in his later approach to autonomy may also be problematic from a neuroscientific perspective. This type of self-control requires a kind of conscious and effortful resistance of our temptations to achieve our goals. I argue that the neuroscience of self-control points out that effortful resistance of temptations is not always the best way to define self-control, and in fact, sometimes temptations can actually help reinforce our long-term goals. I argued that Frankfurt's caring approach, with its essential "active self" providing the motivational authority over our actions, differs from the neuroscientific research of goal conflicts.

I also examined Frankfurt's paradigmatic case of the unwilling drug addict. I considered some of the recent neuroscience of addiction and compared this with Frankfurt's conception of the unwilling addict. I argued that the neuroscience of addiction demonstrates that addiction can cause many pervasive changes to a person's decision-making, self-regulatory, and inhibitory control processes—likely leading to many autonomy issues like risky-decision-making, issues of impulsivity, and generally a reduced ability to inhibit their drug-related activity.

Lastly, I discussed some key differences between how we experience feelings of control and ownership over our actions and Frankfurt's conception of how we experience our desires. I argued that there were two main tensions between Frankfurt's approach and the neuroscience research on this topic. Frankfurt's approach requires a feeling of satisfaction with our desires as a requirement for autonomy. I point out that this may be difficult to achieve based on the prevalence of nonconscious goal-directed behaviour driving much of our everyday behaviours. Additionally, I examined Frankfurt's requirement of relying on these feelings of satisfaction for agential authority. I pointed out that research on the motivation system suggests possible dissociations between feelings of "wanting" and feelings of "liking." Therefore, I argued that these differences between wanting something and liking something may present a problem for Frankfurt's account, which relies on feelings of wanting and liking in determining the legitimacy of our desires.

Given the importance of emotion and motivation in processes like decision-making and control, it might initially seem like Frankfurt's structural theory of autonomy would be more compatible with current neuroscience than Korsgaard's reasons-responsive account. However, I argued that both Korsgaard and Frankfurt have some autonomy requirements that are problematic in light of the neuroscience reviewed. There are key differences between the neuroscience research and how these autonomy theories understand desires as motivating our actions, the role of emotion in decision-making and various forms of control, and relying on our conscious experiences or feelings of agency and control over our actions.

One important question within the autonomy literature is what gives us authority over our actions. Does the way we reason about our desires and motivations grant this authority? Or is it a particular feeling about our desires that makes them autonomous? These questions were explored

in both Korsgaard's and Frankfurt's approaches, where I argued that both accounts of agential authority may not be compatible with the neuroscience reviewed. Korsgaard's requirement of conscious deliberation over our motivations is not supported by the empirical research looking at our motivation system, and how it relates to decision-making, goal-directed behaviours, and control. And, while our affective processes are central to making decisions and control, this is not what Frankfurt means by "feeling satisfied" with our desires. I pointed out that an affective, rewards-based valuation system seems to underlie many of the autonomy related processes, such as decision-making and control. When this system is compromised, we can see behaviours we would term non-autonomous, or less autonomous, such as risky decisions, impulsive behaviours, or a lack of control over one's actions. Hopefully, with a sharper understanding of the neuroscience underlying processes like decision-making, self-control, the motivational system, voluntary action, and our conscious feelings of agency, we can get to a clearer understanding of a highly complex concept like personal autonomy and what it means when this concept is compromised.

Future Directions

One area that was briefly mentioned, but left largely unexplored, is research on the neuroscience of self. I chose to exclude this topic for two main reasons. First, the research on the neuroscience of the self is quite extensive, so it was excluded for matters of scope. Second, I chose to focus primarily on decision-making and control processes because they are fundamental processes for personal autonomy. However, the question of what constitutes the self from a neuroscientific perspective is directly relevant to the question of autonomy and needs further exploration.

For the past couple of decades, there has been extensive research on the neural mechanisms of the self (Northoff et al. 2011; Feinberg 2009; Legrand and Ruby 2009; Northoff and Panksepp 2008; Panksepp 2007; Northoff and Bermpohl 2004; Gallagher and Frith 2003; Damasio 2000). This research indicates that cortical midline structures (e.g., the ventromedial prefrontal cortex, the perigenual anterior cingulate cortex, dorsomedial prefrontal cortex, and the posterior cingulate cortex) are likely involved in self-referential processing (Northoff et al. 2011). Recent research on this subject has also proposed an integrated subcortical-cortical midline system, where interosensorimotor signals (internal stimuli, and other homeostatic and bodily information such as emotional and motivational information) and exterosensorimotor (stimuli external to the body) are integrated as a process-based neural conception of the self (Northoff et al. 2011).

This network of cortical midline structures also seems to strongly overlap with the default-mode network (DMN), characterized as a group of brain regions with a particularly high resting state of intrinsic neural activity when a person is resting (e.g., daydreaming, mind-wandering) or thinking about themselves (Northoff et al. 2011, 57).

Overall, the emerging research on the neural mechanisms of the self has begun to answer some important questions, but it also raises some new avenues to pursue. For instance, it would be an interesting project to explore how this neuroscience of self research relates to the autonomy processes I considered in this dissertation, such as decision-making and self-control.

In particular, it would be relevant to further consider how our values, which are central to decision-making, fit into the kinds of self-referential processing that is being studied in the neuroscience of self. For example, studies on the neuroscience of self have shown that reward-related areas of the brain (e.g., the VTA, vmPFC, and VS) are recruited during the processing of

self-specific stimuli (Northoff and Hayes 2011). This suggests that there may in fact be such a link between reward-related decision-making (key areas related to the determination of our motivational value) and the self. Given that reward-related areas are not just computing values in an impersonal sense but rather are calculating the motivational value as it relates to our self and how it affects our own choice options, this link between reward-processing and self-referential processing is an important avenue to explore. Overall, it would be worth further investigation to see how the neuroscience of self might fit in with a neuroscientific framework for autonomy, and in particular, how the self relates to decision-making.

References

Aarts, Henk, and Ruud Custers. 2012. "Unconscious Goal Pursuit: Nonconscious Goal Regulation and Motivation." In *The Oxford Handbook of Human Motivation*, edited by Richard M. Ryan. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399820.013.0014.

Aarts, Henk, and Ap Dijksterhuis. 2000. "Habits as Knowledge Structures: Automaticity in Goal-Directed Behavior." *Journal of Personality and Social Psychology* (US) 78 (1): 53–63. https://doi.org/10.1037/0022-3514.78.1.53.

Adolphs, Ralph, and David J. Anderson. 2018. *The Neuroscience of Emotion: A New Synthesis*. Princeton university press.

Ainslie, G. W. 1974. "Impulse Control in Pigeons." *Journal of the Experimental Analysis of Behavior* 21 (3): 485–89. https://doi.org/10.1901/jeab.1974.21-485.

Anderson, Michael L. 2014. *After Phrenology: Neural Reuse and the Interactive Brain*. The MIT Press.

Anderson, Michael L., and Luiz Pessoa. 2011. "Quantifying the Diversity of Neural Activations in Individual Brain Regions." *Proceedings of the Annual Meeting of the Cognitive Science Society* 33 (33): 2421–26.

Anselme, Patrick, and Mike J. F. Robinson. 2016. "'Wanting,' 'Liking,' and Their Relation to Consciousness." *Journal of Experimental Psychology: Animal Learning and Cognition* (US) 42 (2): 123–40. https://doi.org/10.1037/xan0000090.

Apps, Matthew A J, Laura L Grima, Sanjay Manohar, and Masud Husain. 2015. "The Role of Cognitive Effort in Subjective Reward Devaluation and Risky Decision-Making." *Scientific Reports* 5 (November): 16880. https://doi.org/10.1038/srep16880.

Aristotle. 2009. *The Basic Works of Aristotle*. With Richard McKeon and C. D. C. Reeve. Random House Publishing Group.

Aron, Adam R. 2007. "The Neural Basis of Inhibition in Cognitive Control." *The Neuroscientist* 13 (3): 214–28. https://doi.org/10.1177/1073858407299288.

Aron, Adam R. 2011. "From Reactive to Proactive and Selective Control: Developing a Richer Model for Stopping Inappropriate Responses." *Biological Psychiatry*, Prefrontal Cortical Circuits Regulating Attention, Behavior and Emotion, vol. 69 (12): e55–68. https://doi.org/10.1016/j.biopsych.2010.07.024.

Aron, Adam R., Trevor W. Robbins, and Russell A. Poldrack. 2004. "Inhibition and the Right Inferior Frontal Cortex." *Trends in Cognitive Sciences* 8 (4): 170–77. https://doi.org/10.1016/j.tics.2004.02.010.

Aron, Adam R., Trevor W. Robbins, and Russell A. Poldrack. 2014. "Inhibition and the Right Inferior Frontal Cortex: One Decade On." *Trends in Cognitive Sciences* 18 (4): 177–85. https://doi.org/10.1016/j.tics.2013.12.003.

Arpaly, Nomy. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford University Press.

Austin, James T., and Jeffrey B. Vancouver. 1996. "Goal Constructs in Psychology: Structure, Process, and Content." *Psychological Bulletin* (US) 120 (3): 338–75. https://doi.org/10.1037/0033-2909.120.3.338.

Banks, William P., and Eve A. Isham. 2011. "Do We Really Know What We Are Doing? Implications of Reported Time of Decision for Theories of Volition." In *Conscious Will*

and Responsibility: A Tribute to Benjamin Libet, edited by Walter Sinnott-Armstrong and Lynn Nadel. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195381641.003.0006.

Bargh, John A. 1990. "Auto-Motives: Preconscious Determinants of Social Interaction." In *Handbook of Motivation and Cognition: Foundations of Social Behavior, Vol. 2.* The Guilford Press.

Bargh, John A. 1997. "The Automaticity of Everyday Life." In *The Automaticity of Everyday Life: Advances in Social Cognition, Vol. 10*, edited by Robert S. Wyer, vol. 10. Advances in Social Cognition. Lawrence Erlbaum Associates Publishers.

Bargh, John A. 2006. "What Have We Been Priming All These Years? On the Development, Mechanisms, and Ecology of Nonconscious Social Behavior." *European Journal of Social Psychology* 36 (2): 147–68. https://doi.org/10.1002/ejsp.336.

Bargh, John A., and Melissa J. Ferguson. 2000. "Beyond Behaviorism: On the Automaticity of Higher Mental Processes." *Psychological Bulletin* (US) 126 (6): 925–45. https://doi.org/10.1037/0033-2909.126.6.925.

Bargh, John A., and Peter M. Gollwitzer. 1994. "Environmental Control of Goal-Directed Action: Automatic and Strategic Contingencies between Situations and Behavior." In *Integrative Views of Motivation, Cognition, and Emotion*. Nebraska Symposium on Motivation, Vol. 41. University of Nebraska Press.

Bargh, John A, Peter M Gollwitzer, Annette Lee-Chai, Kimberly Barndollar, and Roman Trotschel. 2001. "The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals." *Journal of Personality and Social Psychology* 81 (6): 1014–27.

Bargh, John A., Kay L. Schwader, Sarah E. Hailey, Rebecca L. Dyer, and Erica J. Boothby. 2012. "Automaticity in Social-Cognitive Processes." *Trends in Cognitive Sciences* 16 (12): 593–605. https://doi.org/10.1016/j.tics.2012.10.002.

Bari, Andrea, and Trevor W. Robbins. 2013. "Inhibition and Impulsivity: Behavioral and Neural Basis of Response Control." *Progress in Neurobiology* 108: 44–79. https://doi.org/10.1016/j.pneurobio.2013.06.005.

Barrett, Lisa Feldman. 2009. "The Future of Psychology: Connecting Mind to Brain." *Perspectives on Psychological Science* 4 (4): 326–39.

Barrett, Lisa Feldman. 2017a. *How Emotions Are Made: The Secret Life of the Brain*. Mariner Books.

Barrett, Lisa Feldman. 2017b. "The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization." *Social Cognitive and Affective Neuroscience* 12 (1): 1–23. https://doi.org/10.1093/scan/nsx060.

Barrett, Lisa Feldman, and Ajay B. Satpute. 2019. "Historical Pitfalls and New Directions in the Neuroscience of Emotion." *Neuroscience Letters* 693 (February): 9–18. https://doi.org/10.1016/j.neulet.2017.07.045.

Barrett, Lisa Feldman, and Ajay Bhaskar Satpute. 2013. "Large-Scale Brain Networks in Affective and Social Neuroscience: Towards an Integrative Functional Architecture of the Brain." *Current Opinion in Neurobiology* 23 (3): 361–72. https://doi.org/10.1016/j.conb.2012.12.012.

Baumeister, R.F., and T.F. Heatherton. 1996. "Self-Regulation Failure: An Overview." *Psychological Inquiry* 7 (1): 1–15. https://psycnet.apa.org/doi/10.1207/s15327965pli0701_1.

Baumeister, Roy F., Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. 1998. "Ego Depletion: Is the Active Self a Limited Resource?" *Journal of Personality and Social Psychology* (US) 74 (5): 1252–65. https://doi.org/10.1037/0022-3514.74.5.1252.

Baumeister, Roy F., Dianne M. Tice, and Kathleen D. Vohs. 2018. "The Strength Model of Self-Regulation: Conclusions From the Second Decade of Willpower Research." *Perspectives on Psychological Science* 13 (2): 141–45. https://doi.org/10.1177/1745691617716946.

Baxter, Mark G., and Elisabeth A. Murray. 2002. "The Amygdala and Reward." *Nature Reviews Neuroscience* 3 (7): 563–73. https://doi.org/10.1038/nrn875.

Bayne, Tim, and Elisabeth Pacherie. 2007. "Narrators and Comparators: The Architecture of Agentive Self-Awareness." *Synthese* 159 (3): 475–91.

Beam, Elizabeth, Christopher Potts, Russell A. Poldrack, and Amit Etkin. 2021. "A Data-Driven Framework for Mapping Domains of Human Neurobiology." *Nature Neuroscience* 24 (12): 1733–44. https://doi.org/10.1038/s41593-021-00948-9.

Bechara, Antoine, Antonio R. Damasio, Hanna Damasio, and Steven W. Anderson. 1994. "Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex." *Cognition* 50 (1–3): 7–15. https://doi.org/10.1016/0010-0277(94)90018-3.

Bechara, Antoine, Hanna Damasio, and Antonio R. Damasio. 2000. "Emotion, Decision Making and the Orbitofrontal Cortex." *Cerebral Cortex* 10 (3): 295–307. https://doi.org/10.1093/cercor/10.3.295.

Bechara, Antoine, Hanna Damasio, Antonio R. Damasio, and Gregory P. Lee. 1999. "Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision-Making." *The Journal of Neuroscience* 19 (13): 5473–81. https://doi.org/10.1523/JNEUROSCI.19-13-05473.1999.

Bechara, Antoine, Hanna Damasio, Daniel Tranel, and Antonio R Damasio. 1997. "Deciding Advantageously Before Knowing the Advantageous Strategy." *Science* 275 (5304): 1293–95.

Belfiore, Elizabeth. 2006. "Dancing with the Gods: The Myth of the Chariot in Plato's 'Phaedrus.'" *The American Journal of Philology* 127 (2): 185–217.

Bennett, Jonathan. 1974. "The Conscience of Huckleberry Finn." *Philosophy* 49 (188): 123–34.

Berkman, Elliot T. 2018. "The Neuroscience of Self-Control." In *Routledge International Handbook of Self-Control in Health and Well-Being*. Routledge.

Berkman, Elliot T., Emily B. Falk, and Matthew D. Lieberman. 2011. "In the Trenches of Real-World Self-Control: Neural Correlates of Breaking the Link Between Craving and Smoking." *Psychological Science* 22 (4): 498–506.

Berkman, Elliot T., Cendri A. Hutcherson, Jordan L. Livingston, Lauren E. Kahn, and Michael Inzlicht. 2017. "Self-Control as Value-Based Choice." *Current Directions in Psychological Science* 26 (5): 422–28.

Berridge, Kent C. 2004. "Motivation Concepts in Behavioral Neuroscience." *Physiology & Behavior*, Reviews on Ingestive Science, vol. 81 (2): 179–209. https://doi.org/10.1016/j.physbeh.2004.02.004.

Berridge, Kent C. 2007. "The Debate over Dopamine's Role in Reward: The Case for Incentive Salience." *Psychopharmacology* 191 (3): 391–431. https://doi.org/10.1007/s00213-006-0578-x.

Berridge, Kent C, and Terry E Robinson. 1998. "What Is the Role of Dopamine in Reward: Hedonic Impact, Reward Learning, or Incentive Salience?" *Brain Research Reviews* 28 (3): 309–69. https://doi.org/10.1016/S0165-0173(98)00019-8.

Berridge, Kent C., Terry E. Robinson, and Wayne J. Aldridge. 2009. “Dissecting Components of Reward: ‘Liking’, ‘Wanting’, and Learning.” *Current Opinion in Pharmacology* 9: 65–73.

Berridge, Kent C., and Morten L. Kringelbach. 2015. “Pleasure Systems in the Brain.” *Neuron* 86 (3): 646–64. https://doi.org/10.1016/j.neuron.2015.02.018.

Bogacz, Rafal, and Kevin Gurney. 2007. “The Basal Ganglia and Cortex Implement Optimal Decision Making Between Alternative Actions.” *Neural Computation* 19 (2): 442–77. https://doi.org/10.1162/neco.2007.19.2.442.

Bogacz, Rafal, and Tobias Larsen. 2011. “Integration of Reinforcement Learning and Optimal Decision-Making Theories of the Basal Ganglia.” *Neural Computation* 23 (4): 817–51. https://doi.org/10.1162/NECO_a_00103.

Boorman, Erie D., Timothy E. J. Behrens, Mark W. Woolrich, and Matthew F. S. Rushworth. 2009. “How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action.” *Neuron* 62 (5): 733–43. https://doi.org/10.1016/j.neuron.2009.05.014.

Botvinick, Matthew M. 2007. “Conflict Monitoring and Decision Making: Reconciling Two Perspectives on Anterior Cingulate Function.” *Cognitive, Affective, & Behavioral Neuroscience* 7 (4): 356–66. https://doi.org/10.3758/CABN.7.4.356.

Botvinick, Matthew M., Todd S. Braver, Deanna M. Barch, Cameron S. Carter, and Jonathan D. Cohen. 2001. “Conflict Monitoring and Cognitive Control.” *Psychological Review* (US) 108 (3): 624–52. https://doi.org/10.1037/0033-295X.108.3.624.

Botvinick, Matthew M., and Jonathan D. Cohen. 2014. “The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers.” *Cognitive Science* 38 (6): 1249–85. https://doi.org/10.1111/cogs.12126.

Brass, Marcel, and Patrick Haggard. 2007. “To Do or Not to Do: The Neural Signature of Self-Control.” *The Journal of Neuroscience* 27 (34): 9141–45. https://doi.org/10.1523/JNEUROSCI.0924-07.2007.

Brass, Marcel, and Patrick Haggard. 2008. “The What, When, Whether Model of Intentional Action.” *The Neuroscientist* 14 (4): 319–25. https://doi.org/10.1177/1073858408317417.

Braver, Todd S., Marie K. Krug, Kimberly S. Chiew, et al. 2014. “Mechanisms of Motivation-Cognition Interaction: Challenges and Opportunities.” *Cognitive, Affective and Behavioral Neuroscience* (New York, Netherlands) 14 (2): 443–72.

Breiter, Hans C., Itzhak Aharon, Daniel Kahneman, Anders Dale, and Peter Shizgal. 2001. “Functional Imaging of Neural Responses to Expectancy and Experience of Monetary Gains and Losses.” *Neuron* 30 (2): 619–39. https://doi.org/10.1016/s0896-6273(01)00303-8.

Bromberg-Martin, Ethan S., Masayuki Matsumoto, and Okihide Hikosaka. 2010. “Dopamine in Motivational Control: Rewarding, Aversive, and Alerting.” *Neuron* 68 (5): 815–34. https://doi.org/10.1016/j.neuron.2010.11.022.

Bruce, Charles J., and Michael E. Goldberg. 1985. “Primate Frontal Eye Fields. I. Single Neurons Discharging before Saccades.” *Journal of Neurophysiology* 53 (3): 603–35. https://doi.org/10.1152/jn.1985.53.3.603.

Burnod, Yves, Pierre Baraduc, Alexandra Battaglia-Mayer, et al. 1999. “Parieto-Frontal Coding of Reaching: An Integrated Framework.” *Experimental Brain Research* 129 (3): 0325–46. https://doi.org/10.1007/s002210050902.

Buss, Sarah, and Andrea Westlund. 2018. “Personal Autonomy.” In *The Stanford Encyclopedia of Philosophy*, Spring 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

Cai, Xinying, and Camillo Padoa-Schioppa. 2014. “Contributions of Orbitofrontal and Lateral Prefrontal Cortices to Economic Choice and the Good-to-Action Transformation.” *Neuron* 81 (5): 1140–51. https://doi.org/10.1016/j.neuron.2014.01.008.

Carver, Charles S., and Michael F. Scheier. 1998. *On the Self-Regulation of Behavior*. 1st ed. Cambridge University Press.

Caspar, Emilie A., Julia F. Christensen, Axel Cleeremans, and Patrick Haggard. 2016. “Coercion Changes the Sense of Agency in the Human Brain.” *Current Biology* 26 (5): 585–92. https://doi.org/10.1016/j.cub.2015.12.067.

Castro, Daniel C., and Kent C. Berridge. 2014. “Opioid Hedonic Hotspot in Nucleus Accumbens Shell: Mu, Delta, and Kappa Maps for Enhancement of Sweetness ‘Liking’ and ‘Wanting.’” *The Journal of Neuroscience* 34 (12): 4239–50. https://doi.org/10.1523/JNEUROSCI.4458-13.2014.

Cavallo, Justin V., and Gráinne M. Fitzsimons. 2012. “Goal Competition, Conflict, Coordination, and Completion: How Intergoal Dynamics Affect Self-Regulation.” In *Goal-Directed Behavior*, edited by Henk Aarts and Andrew Elliot.

Chen, Mei-Yen, Koji Jimura, Corey N. White, W. Todd Maddox, and Russell A. Poldrack. 2015. “Multiple Brain Networks Contribute to the Acquisition of Bias in Perceptual Decision-Making.” *Frontiers in Neuroscience* 9 (March). https://doi.org/10.3389/fnins.2015.00063.

Chiew, Kimberly S., and Todd S. Braver. 2011. “Positive Affect Versus Reward: Emotional and Motivational Influences on Cognitive Control.” *Frontiers in Psychology* 2 (October): 279. https://doi.org/10.3389/fpsyg.2011.00279.

Childress, Anna Rose, Ronald N. Ehrman, Ze Wang, et al. 2008. “Prelude to Passion: Limbic Activation by ‘Unseen’ Drug and Sexual Cues.” *PLoS ONE* 3 (1): e1506. https://doi.org/10.1371/journal.pone.0001506.

Christman, John. 2003. *Autonomy in Moral and Political Philosophy*. July 28. https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/.

Cisek, Paul. 2007. “Cortical Mechanisms of Action Selection: The Affordance Competition Hypothesis.” *Philosophical Transactions: Biological Sciences* 362 (1485): 1585–99.

Clithero, John A., and Antonio Rangel. 2014. “Informatic Parcellation of the Network Involved in the Computation of Subjective Value.” *Social Cognitive and Affective Neuroscience* 9 (9): 1289–302. https://doi.org/10.1093/scan/nst106.

Cohen, Jonathan D. 2017. “Cognitive Control: Core Constructs and Current Considerations.” In *The Wiley Handbook of Cognitive Control*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118920497.ch1.

Cohon, Rachel. 2004. *Hume’s Moral Philosophy*. October 29. https://plato.stanford.edu/archives/fall2018/entries/hume-moral/.

Cole, Michael W., Jeremy R. Reynolds, Jonathan D. Power, Grega Repovs, Alan Anticevic, and Todd S. Braver. 2013. “Multi-Task Connectivity Reveals Flexible Hubs for Adaptive Task Control.” *Nature Neuroscience* 16 (9): 1348–55.

Conchiglia, Giovannina, Gennaro Della Rocca, and Dario Grossi. 2007. “On a Peculiar Environmental Dependency Syndrome in a Case with Frontal-Temporal Damage: Zelig-like Syndrome.” *Neurocase* 13 (1): 1–5. https://doi.org/10.1080/13554790601160558.

Cosmides, Leda, and John Tooby. 1992. “Cognitive Adaptations for Social Exchange.” In *The Adapted Mind*, edited by Jerome H Barkow, Leda Cosmides, and John Tooby. Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195060232.003.0004.

Cowan, Nelson. 2001. “The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity.” *Behavioral and Brain Sciences* 24 (1): 87–114. https://doi.org/10.1017/S0140525X01003922.

Custers, Ruud, Baruch Eitam, and John A. Bargh. 2012. “Conscious and Unconscious Processes in Goal Pursuit.” In *Goal-Directed Behavior*. Frontiers of Social Psychology. Psychology Press. https://doi.org/10.4324/9780203869666.

Dalley, Jeffrey W., and Trevor W. Robbins. 2017. “Fractionating Impulsivity: Neuropsychiatric Implications.” *Nature Reviews. Neuroscience* (London, United States) 18 (3): 158–71. https://doi.org/10.1038/nrn.2017.8.

Damasio, Antonio R. 1994. *Descartes’ Error: Emotion, Reason and the Human Brain*. Rev. ed. with a new preface. Vintage.

Damasio, Antonio R. 2000. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. 1. Harvest ed. A Harvest Book. Harcourt.

Darwin, Charles. 2014. *On the Origin of Species: By Means of Natural Selection*. Edited by George W. Davidson. Arcturus.

Deci, Edward L., and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. Plenum Press.

Desmurget, Michel, Karen T. Reilly, Nathalie Richard, Alexandru Szathmari, Carmine Mottolese, and Angela Sirigu. 2009. “Movement Intention after Parietal Cortex Stimulation in Humans.” *Science* 324 (5928): 811–13.

Desmurget, Michel, and Angela Sirigu. 2012. “Conscious Motor Intention Emerges in the Inferior Parietal Lobule.” *Current Opinion in Neurobiology*, Decision making, vol. 22 (6): 1004–11. https://doi.org/10.1016/j.conb.2012.06.006.

Dias, R., T. W. Robbins, and A. C. Roberts. 1996. “Dissociation in Prefrontal Cortex of Affective and Attentional Shifts.” *Nature* (London, United States) 380 (6569): 69–72. https://doi.org/10.1038/380069a0.

Dignath, David, Andreas B. Eder, Marco Steinhauser, and Andrea Kiesel. 2020. “Conflict Monitoring and the Affective-Signaling Hypothesis—An Integrative Review.” *Psychonomic Bulletin & Review* 27 (2): 193–216. https://doi.org/10.3758/s13423-019-01668-9.

Döring, Sabine A. 2007. “Seeing What to Do: Affective Perception and Rational Motivation.” *Dialectica* 61 (3): 363–94.

Dreisbach, Gesine. 2006. “How Positive Affect Modulates Cognitive Control: The Costs and Benefits of Reduced Maintenance Capability.” *Brain and Cognition* 60 (1): 11–19. https://doi.org/10.1016/j.bandc.2005.08.003.

Dreisbach, Gesine, and Rico Fischer. 2015. “Conflicts as Aversive Signals for Control Adaptation.” *Current Directions in Psychological Science* 24 (4): 255–60.

Duckworth, Angela L., Tamar Szabó Gendler, and James J. Gross. 2016. “Situational Strategies for Self-Control.” *Perspectives on Psychological Science* 11 (1): 35–55.

Edinger, Ludwig, and H. W. Rand. 1908. “The Relations of Comparative Anatomy to Comparative Psychology.” *Journal of Comparative Neurology and Psychology* 18 (5): 437–57. https://doi.org/10.1002/cne.920180502.

Elliott, Rebecca, Raymond J. Dolan, and Chris D. Frith. 2000. “Dissociable Functions in the Medial and Lateral Orbitofrontal Cortex: Evidence from Human Neuroimaging Studies.” *Cerebral Cortex* 10 (3): 308–17. https://doi.org/10.1093/cercor/10.3.308.

Feinberg, Todd E. 2009. *From Axons to Identity*. Norton Series on Interpersonal Neurobiology Ser, v. 0. W. W. Norton & Company, Incorporated.

Felsen, Gidon, and Peter B. Reiner. 2011. “How the Neuroscience of Decision Making Informs Our Conception of Autonomy.” *AJOB Neuroscience* 2 (3): 3–14. https://doi.org/10.1080/21507740.2011.580489.

Filevich, Elisa, Simone Kühn, and Patrick Haggard. 2012. “Intentional Inhibition in Human Action: The Power of ‘No.’” *Neuroscience & Biobehavioral Reviews* 36 (4): 1107–18. https://doi.org/10.1016/j.neubiorev.2012.01.006.

Fischman, Marian W., and Richard W. Foltin. 1992. “Self-Administration of Cocaine by Humans: A Laboratory Perspective.” In *Cocaine: Scientific and Social Dimensions*. Ciba Foundation Symposium 166. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470514245.ch10.

Fishbach, Ayelet, Ronald S. Friedman, and Arie W. Kruglanski. 2003. “Leading Us Not into Temptation: Momentary Allurements Elicit Overriding Goal Activation.” *Journal of Personality and Social Psychology* (US) 84 (2): 296–309. https://doi.org/10.1037/0022-3514.84.2.296.

Fishbach, Ayelet, Ronald S. Friedman, and Arie W. Kurglanski. 2003. “Leading Us Not Unto Temptation: Momentary Allurements Elicit Overriding Goal Activation.” *Journal of Personality and Social Psychology* 84 (2): 296–309.

FitzGerald, Thomas H. B., Ben Seymour, and Raymond J. Dolan. 2009. “The Role of Human Orbitofrontal Cortex in Value Comparison for Incommensurable Objects.” *The Journal of Neuroscience* 29 (26): 8388–95. https://doi.org/10.1523/JNEUROSCI.0717-09.2009.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Frank, Michael J., and Eric D. Claus. 2006. “Anatomy of a Decision: Striato-Orbitofrontal Interactions in Reinforcement Learning, Decision Making, and Reversal.” *Psychological Review* (US) 113 (2): 300–326. https://doi.org/10.1037/0033-295X.113.2.300.

Frankfurt, Harry G. 1971. “Freedom of the Will and the Concept of a Person.” *The Journal of Philosophy* 68 (1): 5–20. https://doi.org/10.2307/2024717.

Frankfurt, Harry G. 1998a. “Identification and Wholeheartedness.” In *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press. https://doi.org/10.1017/CBO9780511625411.002.

Frankfurt, Harry G. 1998b. “Rationality and the Unthinkable.” In *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Frankfurt, Harry G. 1998c. *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Frankfurt, Harry G. 1999a. “Autonomy, Necessity, and Love.” In *Necessity, Volition, and Love*. Cambridge University Press. https://doi.org/10.1017/CBO9780511624643.012.

Frankfurt, Harry G. 1999b. *Necessity, Volition, and Love*. Cambridge University Press. https://doi.org/10.1017/CBO9780511624643.

Frankfurt, Harry G. 1999c. “The Faintest Passion.” In *Necessity, Volition, and Love*. Cambridge University Press. https://doi.org/10.1017/CBO9780511624643.

Fridman, Joseph, Lisa Feldman Barrett, Jolie B. Wormwood, and Karen S. Quigley. 2019. “Applying the Theory of Constructed Emotion to Police Decision Making.” *Frontiers in Psychology* 10 (September): 1946. https://doi.org/10.3389/fpsyg.2019.01946.

Fried, Itzhak, Patrick Haggard, Biyu J. He, and Aaron Schurger. 2017. “Volition and Action in the Human Brain: Processes, Pathologies, and Reasons.” *The Journal of Neuroscience* 37 (45): 10842–47. https://doi.org/10.1523/JNEUROSCI.2584-17.2017.

Fried, Itzhak, Amiram Katz, Gregory McCarthy, et al. 1991. “Functional Organization of Human Supplementary Motor Cortex Studied by Electrical Stimulation.” *The Journal of Neuroscience* 11 (11): 3656–66. https://doi.org/10.1523/JNEUROSCI.11-11-03656.1991.

Frith, Christopher D., Sarah-Jayne Blakemore, and Daniel M. Wolpert. 2000. “Abnormalities in the Awareness and Control of Action.” *Philosophical Transactions: Biological Sciences* 355 (1404): 1771–88.

Frömer, R., and A. Shenhav. 2022. “Filling the Gaps: Cognitive Control as a Critical Lens for Understanding Mechanisms of Value-Based Decision-Making.” *Neuroscience & Biobehavioral Reviews* 134 (March): 104483. https://doi.org/10.1016/j.neubiorev.2021.12.006.

Fujita, Kentaro. 2011. “On Conceptualizing Self-Control as More Than the Effortful Inhibition of Impulses.” *Personality and Social Psychology Review* 15 (4): 352–66. https://doi.org/10.1177/1088868311411165.

Galla, Brian M., and Angela L. Duckworth. 2015. “More than Resisting Temptation: Beneficial Habits Mediate the Relationship between Self-Control and Positive Life Outcomes.” *Journal of Personality and Social Psychology* (US) 109 (3): 508–25. https://doi.org/10.1037/pspp0000026.

Gallagher, Helen L., and Christopher D. Frith. 2003. “Functional Imaging of ‘Theory of Mind.’” *Trends in Cognitive Sciences* 7 (2): 77–83. https://doi.org/10.1016/S1364-6613(02)00025-6.

Gillan, Claire M., Trevor W. Robbins, Barbara J. Sahakian, Odile A. van den Heuvel, and Guido van Wingen. 2016. “The Role of Habit in Compulsivity.” *European Neuropsychopharmacology*, Compulsivity- a new trans-diagnostic research domain for the Roadmap for Mental Health Research in Europe (ROAMER) and Research Domain Criteria (RDoC) initiatives, vol. 26 (5): 828–40. https://doi.org/10.1016/j.euroneuro.2015.12.033.

Gillebaart, Marleen. 2018. “The ‘Operational’ Definition of Self-Control.” *Frontiers in Psychology* 9 (July): 1231. https://doi.org/10.3389/fpsyg.2018.01231.

Gillebaart, Marleen, and Denise T. D. de Ridder. 2015. “Effortless Self-Control: A Novel Perspective on Response Conflict Strategies in Trait Self-Control.” *Social and Personality Psychology Compass* 9 (2): 88–99. https://doi.org/10.1111/spc3.12160.

Glimcher, Paul W. 2003. “The Neurobiology of Visual-Saccadic Decision Making.” *Annual Review of Neuroscience* (Palo Alto, United States) 26: 133–79.

Glimcher, Paul W., and David L. Sparks. 1992. “Movement Selection in Advance of Action in the Superior Colliculus.” *Nature* (London, United States) 355 (6360): 542–45. https://doi.org/10.1038/355542a0.

Gold, Joshua I., and Michael N. Shadlen. 2007. "The Neural Basis of Decision Making." *Annual Review of Neuroscience* 30 (1): 535–74. https://doi.org/10.1146/annurev.neuro.29.051605.113038.

Gollwitzer, Peter M. 1990. "Action Phases and Mind-Sets." In *Handbook of Motivation and Cognition: Foundations of Social Behavior, Vol. 2.* The Guilford Press.

Gremel, Christina M., and Rui M. Costa. 2013. "Orbitofrontal and Striatal Circuits Dynamically Encode the Shift between Goal-Directed and Habitual Actions." *Nature Communications* 4: 2264. https://doi.org/10.1038/ncomms3264.

Grèzes, Julie, and Jean Decety. 2001. "Functional Anatomy of Execution, Mental Simulation, Observation, and Verb Generation of Actions: A Meta-analysis." *Human Brain Mapping* 12 (1): 1–19. https://doi.org/10.1002/1097-0193(200101)12:1<1::AID-HBM10>3.0.CO;2-V.

Haber, Suzanne N. 2011. "Neuroanatomy of Reward: A View from the Ventral Striatum." In *Neurobiology of Sensation and Reward*, edited by Jay A. Gottfried. Frontiers in Neuroscience. CRC Press/Taylor & Francis. http://www.ncbi.nlm.nih.gov/books/NBK92777/.

Haber, Suzanne N. 2014. "The Place of Dopamine in the Cortico-Basal Ganglia Circuit." *Neuroscience* 282 (December): 248–57. https://doi.org/10.1016/j.neuroscience.2014.10.008.

Haber, Suzanne N. 2016. "Corticostriatal Circuitry." *Dialogues in Clinical Neuroscience* 18 (1): 7–21. https://doi.org/10.31887/DCNS.2016.18.1/shaber.

Haggard, Patrick. 2008. "Human Volition: Towards a Neuroscience of Will." *Nature Reviews Neuroscience* 9 (12): 934–46. https://doi.org/10.1038/nrn2497.

Haggard, Patrick. 2017. "Sense of Agency in the Human Brain." *Nature Reviews Neuroscience* 18 (4): 196–207. https://doi.org/10.1038/nrn.2017.14.

Haggard, Patrick, and Martin Eimer. 1999. "On the Relation between Brain Potentials and the Awareness of Voluntary Movements." *Experimental Brain Research* 126 (1): 128–33. https://doi.org/10.1007/s002210050722.

Hanes, Doug P., and Jeffrey D. Schall. 1996. "Neural Control of Voluntary Movement Initiation." *Science* 274 (5286): 427–30. https://doi.org/10.1126/science.274.5286.427.

Hanes, Doug P., Kirk G. Thompson, and JeffreyD. Schall. 1995. "Relationship of Presaccadic Activity in Frontal Eye Field and Supplementary Eye Field to Saccade Initiation in Macaque: Poisson Spike Train Analysis." *Experimental Brain Research* 103 (1). https://doi.org/10.1007/BF00241967.

Hare, Todd A., Jonathan Malmaud, and Antonio Rangel. 2011. "Focusing Attention on the Health Aspects of Foods Changes Value Signals in vmPFC and Improves Dietary Choice." *Journal of Neuroscience* 31 (30): 11077–87.

Hariri, Ahmad R., Susan Y. Bookheimer, and John C. Mazziotta. 2000. "Modulating Emotional Responses: Effects of a Neocortical Network on the Limbic System." *NeuroReport* 11 (1): 43–48. https://doi.org/10.1097/00001756-200001170-00009.

Hassan, Anhar, and Keith A. Josephs. 2016. "Alien Hand Syndrome." Topical Collection on Behavior. *Current Neurology and Neuroscience Reports* (Philadelphia, Netherlands) 16 (8): 1–10. https://doi.org/10.1007/s11910-016-0676-z.

Heatherton, Todd F., and Dylan D. Wagner. 2011. "Cognitive Neuroscience of Self-Regulation Failure." *Trends in Cognitive Sciences* 15 (3): 132–39. https://doi.org/10.1016/j.tics.2010.12.005.

Hennecke, Marie, Thomas Czikmantori, and Veronika Brandstätter. 2019. “Doing Despite Disliking: Self-Regulatory Strategies in Everyday Aversive Activities.” *European Journal of Personality* 33 (1): 104–28. https://doi.org/10.1002/per.2182.

Hikosaka, Okihide, Yoriko Takikawa, and Reiko Kawagoe. 2000. “Role of the Basal Ganglia in the Control of Purposive Saccadic Eye Movements.” *Physiological Reviews* 80 (3): 953–953.

Hoemann, Katie, and Lisa Feldman Barrett. 2019. “Concepts Dissolve Artificial Boundaries in the Study of Emotion and Cognition, Uniting Body, Brain, and Mind.” *Cognition and Emotion* 33 (1): 67–76. https://doi.org/10.1080/02699931.2018.1535428.

Hofmann, Wilhelm, Malte Friese, and Fritz Strack. 2009. “Impulse and Self-Control from a Dual-Systems Perspective.” *Perspectives on Psychological Science* 4 (2): 162–76.

Holland, Peter C, and Michela Gallagher. 2004. “Amygdala–Frontal Interactions and Reward Expectancy.” *Current Opinion in Neurobiology* 14 (2): 148–55. https://doi.org/10.1016/j.conb.2004.03.007.

Holroyd, Clay B., and Michael G. H. Coles. 2002. “The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity.” *Psychological Review* (US) 109 (4): 679–709. https://doi.org/10.1037/0033-295X.109.4.679.

Horvitz, J. C. 2000. “Mesolimbocortical and Nigrostriatal Dopamine Responses to Salient Non-Reward Events.” *Neuroscience* 96 (4): 651–56. https://doi.org/10.1016/S0306-4522(00)00019-1.

Huang, Julie Y., and John A. Bargh. 2014. “The Selfish Goal: Autonomously Operating Motivational Structures as the Proximate Cause of Human Judgment and Behavior.” Target Article. *Behavioral and Brain Sciences* (New York, United Kingdom) 37 (2): 121–35. https://doi.org/10.1017/S0140525X13000290.

Hubel, D H, and T N Wiesel. 1959. “RECEPTIVE FIELDS OF SINGLE NEURONES IN THE CAT’S STRIATE CORTEX.” *The Journal of Physiology* 148 (3): 574–91.

Hume, David. 1985. *A Treatise of Human Nature*. Reprinted. Edited by Ernest Campbell Mossner. Penguin Classics. Penguin Books.

Hunt, Laurence T., Nils Kolling, Alireza Soltani, Mark W. Woolrich, Matthew F. S. Rushworth, and Timothy E. J. Behrens. 2012. “Mechanisms Underlying Cortical Activity during Value-Guided Choice.” *Nature Neuroscience* (New York, United States) 15 (3): 470–76, S1-3. https://doi.org/10.1038/nn.3017.

Hutcherson, Cendri A., Benjamin Bushong, and Antonio Rangel. 2015. “A Neurocomputational Model of Altruistic Choice and Its Implications.” *Neuron* 87 (2): 451–62. https://doi.org/10.1016/j.neuron.2015.06.031.

Hyman, Steven E. 2010. “The Neurobiology of Addition: Implications for the Voluntary Control of Behavior.” In *Neuroethics: An Introduction with Readings*. MIT Press.

Hyman, Steven E., Robert C. Malenka, and Eric J. Nestler. 2006. “NEURAL MECHANISMS OF ADDICTION: The Role of Reward-Related Learning and Memory.” *Annual Review of Neuroscience* 29: 565–98. https://doi.org/10.1146/annurev.neuro.29.051605.113009.

Iaccarino, Leonardo, Sergio Chieffi, and Alessandro Iavarone. 2014. “Utilization Behavior: What Is Known and What Has to Be Known?” *Behavioural Neurology* (Cairo, United Kingdom) 2014: 1–9. https://doi.org/10.1155/2014/297128.

Ikemoto, Satoshi, Chen Yang, and Aaron Tan. 2015. “Basal Ganglia Circuit Loops, Dopamine and Motivation: A Review and Enquiry.” *Behavioural Brain Research* 290 (September): 17–31. https://doi.org/10.1016/j.bbr.2015.04.018.

Inzlicht, Michael, Bruce D. Bartholow, and Jacob B. Hirsh. 2015. “Emotional Foundations of Cognitive Control.” *Trends in Cognitive Sciences* 19 (3): 126–32. https://doi.org/10.1016/j.tics.2015.01.004.

Inzlicht, Michael, Kaitlyn M Werner, Julia L Briskin, and Brent W Roberts. 2021. “Integrating Models of Self-Regulation.” *Annual Review of Psychology* 72: 319–45.

Jackson, J Hughlings. 1884. “The Croonian Lectures on Evolution and Dissolution of the Nervous System.” *The British Medical Journal* 1 (1214): 703–7.

Jones, Karen. 2003. “Emotion, Weakness of WIll, and the Normative Conception of Agency.” In *Philosophy and the Emotions*, edited by Anthony Hatzimoysis. Cambridge University Press.

Joseph, Michael H, Krishna Datla, and Andrew M. J Young. 2003. “The Interpretation of the Measurement of Nucleus Accumbens Dopamine by in Vivo Dialysis: The Kick, the Craving or the Cognition?” *Neuroscience & Biobehavioral Reviews* 27 (6): 527–41. https://doi.org/10.1016/j.neubiorev.2003.09.001.

Kable, Joseph W., and Paul W. Glimcher. 2007. “The Neural Correlates of Subjective Value during Intertemporal Choice.” *Nature Neuroscience* (New York, United States) 10 (12): 1625–33. https://doi.org/10.1038/nn2007.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Doubleday Canada. http://ebookcentral.proquest.com/lib/waterloo/detail.action?docID=6813985.

Kant, Immanuel. 1998. *Groundwork of the Metaphysics of Morals*. Edited by Mary J. Gregor. Cambridge Texts in the History of Philosophy. Cambridge University Press.

Kapur, Shitij. 2003. “Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia.” *American Journal of Psychiatry* 160 (1): 13–23. https://doi.org/10.1176/appi.ajp.160.1.13.

Kelley, Ann E., and Kent C. Berridge. 2002. “The Neuroscience of Natural Rewards: Relevance to Addictive Drugs.” *The Journal of Neuroscience* 22 (9): 3306–11. https://doi.org/10.1523/JNEUROSCI.22-09-03306.2002.

Kelley, William M., Dylan D. Wagner, and Todd F. Heatherton. 2015. “In Search of a Human Self-Regulation System.” *Annual Review of Neuroscience* 38 (1): 389–411. https://doi.org/10.1146/annurev-neuro-071013-014243.

Kennerley, Steven W., and Jonathan D. Wallis. 2009. “Encoding of Reward and Space During a Working Memory Task in the Orbitofrontal Cortex and Anterior Cingulate Sulcus.” *Journal of Neurophysiology* 102 (6): 3352–64. https://doi.org/10.1152/jn.00273.2009.

Keren, Gideon, and Yaacov Schul. 2009. “Two Is Not Always Better Than One: A Critical Evaluation of Two-System Theories.” *Perspectives on Psychological Science* 4 (6): 533–50.

Koob, George F., and Nora D. Volkow. 2010. “Neurocircuitry of Addiction.” *Neuropsychopharmacology* (New York, United States) 35 (1): 217–38. https://doi.org/10.1038/npp.2009.110.

Kool, W., and Matthew M. Botvinick. 2014. “A Labor/Leisure Tradeoff in Cognitive Control.” *Journal of Experimental Psychology: General* 143: 131–41.

Kool, Wouter, Joseph T. McGuire, Zev B. Rosen, and Matthew M. Botvinick. 2010. "Decision Making and the Avoidance of Cognitive Demand." *Journal of Experimental Psychology: General* 139 (4): 665–82. https://doi.org/10.1037/a0020198.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge University Press.

Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press.

Krain, Amy L., Amanda M. Wilson, Robert Arbuckle, F. Xavier Castellanos, and Michael P. Milham. 2006. "Distinct Neural Mechanisms of Risk and Ambiguity: A Meta-Analysis of Decision-Making." *NeuroImage* (Amsterdam, United Kingdom) 32 (1): 477–84. https://doi.org/10.1016/j.neuroimage.2006.02.047.

Krönke, Klaus-Martin, Max Wolff, Holger Mohr, et al. 2020. "Predicting Real-Life Self-Control From Brain Activity Encoding the Value of Anticipated Future Outcomes." *Psychological Science* 31 (3): 268–79. https://doi.org/10.1177/0956797619896357.

Kruglanski, Arie W, James Y Shah, Ayelet Fishbach, Ron Friedman, Woo Young Chun, and David Sleeth-Keppler. 2002. "A Theory of Goal Systems." *Advances in Experimental Social Psychology* 34: 331–78.

Kuffler, Stephen W. 1953. "DISCHARGE PATTERNS AND FUNCTIONAL ORGANIZATION OF MAMMALIAN RETINA." *Journal of Neurophysiology* 16 (1): 37–68. https://doi.org/10.1152/jn.1953.16.1.37.

Legrand, Dorothée, and Perrine Ruby. 2009. "What Is Self-Specific? Theoretical Investigation and Critical Review of Neuroimaging Results." *Psychological Review* (US) 116 (1): 252–82. https://doi.org/10.1037/a0014172.

Lepora, Nathan F., and Kevin N. Gurney. 2012. "The Basal Ganglia Optimize Decision Making over General Perceptual Hypotheses." *Neural Computation* 24 (11): 2924–45. https://doi.org/10.1162/NECO_a_00360.

Lerner, Jennifer S., Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. "Emotion and Decision Making." *Annual Review of Psychology* 66 (1): 799–823. https://doi.org/10.1146/annurev-psych-010213-115043.

Levy, Neil. 2015. "Addiction, Autonomy, and Informed Consent: On and Off the Garden Path." *Journal of Medicine and Philosophy*, December 13, jhv033. https://doi.org/10.1093/jmp/jhv033.

Lhermitte, F., B. Pillon, and M. Serdaru. 1986. "Human Autonomy and the Frontal Lobes. Part I: Imitation and Utilization Behavior: A Neuropsychological Study of 75 Patients." *Annals of Neurology* 19 (4): 326–34. https://doi.org/10.1002/ana.410190404.

Libet, Benjamin. 1985. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8 (4): 529–39. https://doi.org/10.1017/S0140525X00044903.

Libet, Benjamin, Elwood W Wright, Curtis A Gleason, and Dennis K. Pearl. 1983. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential)." *Brain* 106: 623–42.

Lieberman, Matthew D., Naomi I. Eisenberger, Molly J. Crockett, Sabrina M. Tom, Jennifer H. Pfeifer, and Baldwin M. Way. 2007. "Putting Feelings into Words: Affect Labeling Disrupts Amygdala Activity in Response to Affective Stimuli." *Psychological Science* 18 (5): 421–28.

Lindquist, Kristen A., and Lisa Feldman Barrett. 2012. “A Functional Architecture of the Human Brain: Emerging Insights from the Science of Emotion.” *Trends in Cognitive Sciences* 16 (11): 533–40. https://doi.org/10.1016/j.tics.2012.09.005.

Lindquist, Kristen A., Tor D. Wager, Hedy Kober, Eliza Bliss-Boreau, and Lisa Feldman Barrett. 2012. “The Brain Basis of Emotion: A Meta-Analytic Review.” *Behavioral and Brain Sciences* (New York, United Kingdom) 35 (3): 121–43. https://doi.org/10.1017/S0140525X11000446.

Lindström, Björn R., Isak Berglund Mattsson-Mårn, Armita Golkar, and Andreas Olsson. 2013. “In Your Face: Risk of Punishment Enhances Cognitive Control and Error-Related Activity in the Corrugator Supercilii Muscle.” *Public Library of Science One* 8 (6): e65692.

Litt, Abninder, Chris Eliasmith, and Paul Thagard. 2008. “Neural Affective Decision Theory: Choices, Brains, and Emotions.” *Cognitive Systems Research* 9 (4): 252–73. https://doi.org/10.1016/j.cogsys.2007.11.001.

Locke, Edwin A., and Gary P. Latham. 2002. “Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey.” *American Psychologist* (US) 57 (9): 705–17. https://doi.org/10.1037/0003-066X.57.9.705.

London, Edythe D., Monique Ernst, Steven Grant, Katherine Bonson, and Aviv Weinstein. 2000. “Orbitofrontal Cortex and Human Drug Abuse: Functional Imaging.” *Cerebral Cortex* 10 (3): 334–42. https://doi.org/10.1093/cercor/10.3.334.

Lopez, Richard B., Wilhelm Hofmann, Dylan D. Wagner, William M. Kelley, and Todd F. Heatherton. 2014. “Neural Predictors of Giving in to Temptation in Daily Life.” *Psychological Science* 25 (7): 1337–44.

Lou, Bin, Wha-Yin Hsu, and Paul Sajda. 2015. “Perceptual Salience and Reward Both Influence Feedback-Related Neural Activity Arising from Choice.” *The Journal of Neuroscience* 35 (38): 13064–75. https://doi.org/10.1523/JNEUROSCI.1601-15.2015.

Loukas, Constantinos, and Peter Brown. 2004. “Online Prediction of Self-Paced Hand-Movements from Subthalamic Activity Using Neural Networks in Parkinson’s Disease.” *Journal of Neuroscience Methods* 137 (2): 193–205. https://doi.org/10.1016/j.jneumeth.2004.02.017.

Lynn, Margaret T., Paul S. Muhle-Karbe, and Marcel Brass. 2014. “Controlling the Self: The Role of the Dorsal Frontomedian Cortex in Intentional Inhibition.” *Neuropsychologia* 65 (December): 247–54. https://doi.org/10.1016/j.neuropsychologia.2014.09.009.

MacLean, Paul D. 1990. *The Triune Brain in Evolution: Role in Paleocerebral Functions*. Plenum Press.

Mannella, Francesco, Kevin Gurney, and Gianluca Baldassarre. 2013. “The Nucleus Accumbens as a Nexus between Values and Goals in Goal-Directed Behavior: A Review and a New Hypothesis.” *Frontiers in Behavioral Neuroscience* 7 (October). https://doi.org/10.3389/fnbeh.2013.00135.

McClure, Samuel M., Nathaniel D. Daw, and P. Read Montague. 2003. “A Computational Substrate for Incentive Salience.” *Trends in Neurosciences* 26 (8): 423–28. https://doi.org/10.1016/S0166-2236(03)00177-2.

McIntyre, Alison, and Amélie Oksenberg Rorty. 1990. “Is Akratic Action Always Irrational?” In *Identity, Character, and Morality*, edited by Am Oksenberg Rorty and Owen Flanagan. The MIT Press.

Meder, David, Damian Marc Herz, James Benedict Rowe, Stéphane Lehéricy, and Hartwig Roman Siebner. 2019. "The Role of Dopamine in the Brain - Lessons Learned from Parkinson's Disease." *NeuroImage* (Amsterdam, United Kingdom) 190 (April): 79–93. https://doi.org/10.1016/j.neuroimage.2018.11.021.

Mele, Alfred R. 2007. "Decisions, Intentions, Urges, and Free Will: Why Libet Has Not Shown What He Says He Has." In *Causation and Explanation*, edited by Joseph Keim Campbell, Michael O'Rourke, and Harry S. Silverstein. The MIT Press. https://doi.org/10.7551/mitpress/1753.003.0014.

Metcalfe, Janet, and Walter Mischel. 1999. "A Hot/Cool-System Analysis of Delay of Gratification: Dynamics of Willpower." *Psychological Review* (US) 106 (1): 3–19. https://doi.org/10.1037/0033-295X.106.1.3.

Milyavskaya, Marina, Elliot T. Berkman, and Denise T. D. De Ridder. 2019. "The Many Faces of Self-Control: Tacit Assumptions and Recommendations to Deal with Them." *Motivation Science* 5 (1): 79–85. https://doi.org/10.1037/mot0000108.

Mischel, Walter. 1973. "Toward a Cognitive Social Learning Reconceptualization of Personality." *Psychological Review* (US) 80 (4): 252–83. https://doi.org/10.1037/h0035002.

Mischel, Walter. 1974. "Processes in Delay of Gratification." *Advances in Experimental Social Psychology* 7: 249–92. https://doi.org/10.1016/S0065-2601(08)60039-8.

Mischel, Walter, Nancy Cantor, and Scott Feldman. 1996. "Principles of Self-Regulation: The Nature of Willpower and Self-Control." In *Social Psychology: Handbook of Basic Principles*, edited by Arie W. Kruglanski, Kruglanski, Arie W., and Edward T. Higgins. The Guilford Press.

Mischel, Walter, Ebbe B. Ebbesen, and Antonette Raskoff Zeiss. 1972. "Cognitive and Attentional Mechanisms in Delay of Gratification." *Journal of Personality and Social Psychology* (US) 21 (2): 204–18. https://doi.org/10.1037/h0032198.

Mishkin, Mortimer. 1964. "Perseveration of Central Sets after Frontal Lesions in Monkeys." In *The Frontal Granular Cortex and Behavior*, edited by J. M. Warren and K. Akerts. McGraw-Hill.

Miyake, Akira, Naomi P. Friedman, Michael J. Emerson, Alexander H. Witzki, Amy Howerter, and Tor D. Wager. 2000. "The Unity and Diversity of Executive Functions and Their Contributions to Complex 'Frontal Lobe' Tasks: A Latent Variable Analysis." *Cognitive Psychology* 41 (1): 49–100. https://doi.org/10.1006/cogp.1999.0734.

Mogenson, G, D Jones, and C Yim. 1980. "From Motivation to Action: Functional Interface between the Limbic System and the Motor System." *Progress in Neurobiology* 14 (2–3): 69–97. https://doi.org/10.1016/0301-0082(80)90018-0.

Montague, P. Read, Steven E. Hyman, and Jonathan D. Cohen. 2004. "Computational Roles for Dopamine in Behavioural Control." Insight Review Articles. *Nature* (London, United States) 431 (7010): 760–67. https://doi.org/10.1038/nature03015.

Morrison, Sara E., Vincent B. McGinty, Johann Du Hoffmann, and Saleem M. Nicola. 2017. "Limbic-Motor Integration by Neural Excitations and Inhibitions in the Nucleus Accumbens." *Journal of Neurophysiology* 118 (5): 2549–67. https://doi.org/10.1152/jn.00465.2017.

Moskowitz, Gordon B. 2010. "On the Control Over Stereotype Activation and Stereotype Inhibition." *Social and Personality Psychology Compass* 4 (2): 140–58. https://doi.org/10.1111/j.1751-9004.2009.00251.x.

Mountcastle, Vernon B. 1957. “MODALITY AND TOPOGRAPHIC PROPERTIES OF SINGLE NEURONS OF CAT’S SOMATIC SENSORY CORTEX.” *Journal of Neurophysiology* 20 (4): 408–34. https://doi.org/10.1152/jn.1957.20.4.408.

Naqvi, Nasir H., and Antoine Bechara. 2009. “The Hidden Island of Addiction: The Insula.” *Trends in Neurosciences* 32 (1): 56–67. https://doi.org/10.1016/j.tins.2008.09.009.

Newsome, William T., Kenneth H. Britten, and J. Anthony Movshon. 1989. “Neuronal Correlates of a Perceptual Decision.” *Nature* 341 (6237): 52–54. https://doi.org/10.1038/341052a0.

Nigg, Joel T. 2017. “Annual Research Review: On the Relations among Self-Regulation, Self-Control, Executive Functioning, Effortful Control, Cognitive Control, Impulsivity, Risk-Taking, and Inhibition for Developmental Psychopathology.” *Journal of Child Psychology and Psychiatry* 58 (4): 361–83. https://doi.org/10.1111/jcpp.12675.

Noonan, M. P., J. Sallet, P. H. Rudebeck, M. J. Buckley, and M. F. Rushworth. 2010. “Does the Medial Orbitofrontal Cortex Have a Role in Social Valuation?” *European Journal of Neuroscience* 31 (12): 2341–51. https://doi.org/10.1111/j.1460-9568.2010.07271.x.

Northoff, Georg. 2006. “Is Emotion Regulation Self-Regulation?” *Trends in Cognitive Sciences* 9 (9): 408–9. https://doi.org/10.1016/j.tics.2005.06.005.

Northoff, Georg. 2017. “Personal Identity and Cortical Midline Structure (CMS): Do Temporal Features of CMS Neural Activity Transform Into ‘Self-Continuity’?” *Psychological Inquiry* 28 (2/3): 122–31.

Northoff, Georg, and Felix Bermpohl. 2004. “Cortical Midline Structures and the Self.” *Trends in Cognitive Sciences* 8 (3): 102–7. https://doi.org/10.1016/j.tics.2004.01.004.

Northoff, Georg, and Dave J. Hayes. 2011. “Is Our Self Nothing but Reward?” *Biological Psychiatry*, Nucleus Accumbens Neuroadaptations and Relapse in Addiction, vol. 69 (11): 1019–25. https://doi.org/10.1016/j.biopsych.2010.12.014.

Northoff, Georg, and Jaak Panksepp. 2008. “The Trans-Species Concept of Self and the Subcortical–Cortical Midline System.” *Trends in Cognitive Sciences* 12 (7): 259–64. https://doi.org/10.1016/j.tics.2008.04.007.

Northoff, Georg, Pengmin Qin, and Todd E. Feinberg. 2011. “Brain Imaging of the Self – Conceptual, Anatomical and Methodological Issues.” *Consciousness and Cognition*, Brain and Self: Bridging the Gap, vol. 20 (1): 52–63. https://doi.org/10.1016/j.concog.2010.09.011.

Ochsner, Kevin N., Rebecca D. Ray, Jeffrey C. Cooper, et al. 2004. “For Better or for Worse: Neural Systems Supporting the Cognitive down- and up-Regulation of Negative Emotion.” *NeuroImage* (Amsterdam, United Kingdom) 23 (2): 483–99. https://doi.org/10.1016/j.neuroimage.2004.06.030.

O’Doherty, J, E T Rolls, S Francis, et al. 2000. “Sensory-Specific Satiety-Related Olfactory Activation of the Human Orbitofrontal Cortex.” *Neuroreport* 11 (4): 893–97.

Okon-Singer, Hadas, Talma Hendler, Luiz Pessoa, and Alexander J. Shackman. 2015. “The Neurobiology of Emotion–Cognition Interactions: Fundamental Questions and Strategies for Future Research.” *Frontiers in Human Neuroscience* 9 (February): 1–14. https://doi.org/10.3389/fnhum.2015.00058.

O’Reilly, Randall C., and Michael J. Frank. 2006. “Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia.” *Neural Computation* 18 (2): 283–328. https://doi.org/10.1162/089976606775093909.

Owen, Adrian M., Kathryn M. McMillan, Angela R. Laird, and Ed Bullmore. 2005. "N-back Working Memory Paradigm: A Meta-analysis of Normative Functional Neuroimaging Studies." *Human Brain Mapping* 25 (1): 46–59. https://doi.org/10.1002/hbm.20131.

Packard, Mark G., and Barbara J. Knowlton. 2002. "Learning and Memory Functions of the Basal Ganglia." *Annual Review of Neuroscience* 25 (Volume 25, 2002): 563–93. https://doi.org/10.1146/annurev.neuro.25.112701.142937.

Panksepp, Jaak. 2007. "The Neuroevolutionary and Neuroaffective Psychobiology of the Prosocial Brain." In *Oxford Handbook of Evolutionary Psychology*, edited by Louise Barrett and Robin Dunbar. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568308.013.0012.

Papez, James W. 1937. "A Proposed Mechanism of Emotion." *Archives of Neurology & Psychiatry* 38 (4): 725–43. https://doi.org/10.1176/jnp.7.1.103.

Pessiglione, Mathias, Liane Schmidt, Bogdan Draganski, et al. 2007. "How the Brain Translates Money into Force: A Neuroimaging Study of Subliminal Motivation." *Science* 316 (5826): 904–6.

Pessoa, Luiz. 2008. "On the Relationship between Emotion and Cognition." *Nature Reviews Neuroscience* 9: 148–58.

Pessoa, Luiz. 2009. "How Do Emotion and Motivation Direct Executive Control?" *Trends in Cognitive Sciences* 13 (4): 160–66. https://doi.org/10.1016/j.tics.2009.01.006.

Pessoa, Luiz. 2013. *The Cognitive-Emotional Brain: From Interactions to Integration*. MIT press.

Pessoa, Luiz. 2018. "Understanding Emotion with Brain Networks." *Current Opinion in Behavioral Sciences*, Emotion-cognition interactions, vol. 19 (February): 19–25. https://doi.org/10.1016/j.cobeha.2017.09.005.

Pessoa, Luiz. 2022. *The Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together*. The MIT Press. https://doi.org/10.7551/mitpress/14636.001.0001.

Pessoa, Luiz, Loreta Medina, and Ester Desfilis. 2021. "Refocusing Neuroscience: Moving Away from Mental Categories and towards Complex Behaviours." *Philosophical Transactions of the Royal Society B: Biological Sciences* 377 (1844): 20200534. https://doi.org/10.1098/rstb.2020.0534.

Peterson, Andrew. 2019. "Should Neuroscience Inform Judgements of Decision-Making Capacity?" *Neuroethics* 12 (2): 133–51. https://doi.org/10.1007/s12152-018-9369-4.

Phan, K. Luan, Daniel A. Fitzgerald, Pradeep J. Nathan, Gregory J. Moore, Thomas W. Uhde, and Manuel E. Tancer. 2005. "Neural Substrates for Voluntary Suppression of Negative Affect: A Functional Magnetic Resonance Imaging Study." *Biological Psychiatry* 57 (3): 210–19. https://doi.org/10.1016/j.biopsych.2004.10.030.

Pickens, Charles L., Michael P. Saddoris, Barry Setlow, Michela Gallagher, Peter C. Holland, and Geoffrey Schoenbaum. 2003. "Different Roles for Orbitofrontal Cortex and Basolateral Amygdala in a Reinforcer Devaluation Task." *The Journal of Neuroscience* 23 (35): 11078–84. https://doi.org/10.1523/JNEUROSCI.23-35-11078.2003.

Plato. 1997. *Complete Works*. Edited by John M. Cooper and D. S. Hutchinson. Hackett Pub.

Platt, Michael L., and Paul W. Glimcher. 1999. "Neural Correlates of Decision Variables in Parietal Cortex." *Nature* 400 (6741): 233–38. https://doi.org/10.1038/22268.

Poldrack, Russell A. 2006. "Can Cognitive Processes Be Inferred from Neuroimaging Data?" *Trends in Cognitive Sciences* 10 (2): 59–63. https://doi.org/10.1016/j.tics.2005.12.004.

Poldrack, Russell A. 2018. *The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts*. First paperback printing. Princeton University Press.

Poldrack, Russell A., Yaroslav O. Halchenko, and Stephen José Hanson. 2009. "Decoding the Large-Scale Structure of Brain Function by Classifying Mental States Across Individuals." *Psychological Science* 20 (11): 1364–72.

Price, Cathy J., and Karl J. Friston. 2005. "Functional Ontologies for Cognition: The Systematic Definition of Structure and Function." *Cognitive Neuropsychology* 22 (3–4): 262–75. https://doi.org/10.1080/02643290442000095.

Rahman, Sufia S, Kevin Mulvihill, Christian M Wood, et al. 2021. "Differential Contribution of Anterior and Posterior Midcingulate Subregions to Distal and Proximal Threat Reactivity in Marmosets." *Cerebral Cortex* 31 (10): 4765–80. https://doi.org/10.1093/cercor/bhab121.

Redgrave, P., T. J. Prescott, and K. Gurney. 1999. "The Basal Ganglia: A Vertebrate Solution to the Selection Problem?" *Neuroscience* 89 (4): 1009–23. https://doi.org/10.1016/S0306-4522(98)00319-4.

Ridder, Denise T. D. de, Benjamin J. de Boer, Peter Lugtig, Arnold B. Bakker, and Edwin A. J. van Hooft. 2011. "Not Doing Bad Things Is Not Equivalent to Doing the Right Thing: Distinguishing between Inhibitory and Initiatory Self-Control." *Personality and Individual Differences*, Special Issue on Anxiety (dedicated to the memory of Professor Blazej Szymura), vol. 50 (7): 1006–11. https://doi.org/10.1016/j.paid.2011.01.015.

Roberts, A.C., and J.D. Wallis. 2000. "Inhibitory Control and Affective Processing in the Prefrontal Cortex: Neuropsychological Studies in the Common Marmoset." *Cerebral Cortex* 10 (3): 252–62.

Robinson, Mike J. F., Patrick Anselme, Kristen Suchomel, and Kent C. Berridge. 2015. "Amphetamine-Induced Sensitization and Reward Uncertainty Similarly Enhance Incentive Salience for Conditioned Cues." *Behavioral Neuroscience* (US) 129 (4): 502–11. https://doi.org/10.1037/bne0000064.

Robinson, Terry E, and Kent C. Berridge. 1993. "The Neural Basis of Drug Craving: An Incentive-Sensitization Theory of Addiction." *Brain Research Reviews* 18 (3): 247–91.

Roesch, Matthew R., and Carl R. Olson. 2004. "Neuronal Activity Related to Reward Value and Motivation in Primate Frontal Cortex." *Science* 304 (5668): 307–10.

Rolls, Edmund T. 2000. "The Orbitofrontal Cortex and Reward." *Cerebral Cortex* 10 (3): 284–94. https://doi.org/10.1093/cercor/10.3.284.

Rolls, Edmund T. 2004. "The Functions of the Orbitofrontal Cortex." *Brain and Cognition*, Development of Orbitofrontal Function, vol. 55 (1): 11–29. https://doi.org/10.1016/S0278-2626(03)00277-X.

Roskies, Adina L. 2010. "How Does Neuroscience Affect Our Conception of Volition?" *Annual Review of Neuroscience* 33 (Volume 33, 2010): 109–30. https://doi.org/10.1146/annurev-neuro-060909-153151.

Rudebeck, Peter H., and Elisabeth A. Murray. 2011. "Dissociable Effects of Subtotal Lesions within the Macaque Orbital Prefrontal Cortex on Reward-Guided Behavior." *The Journal of Neuroscience* 31 (29): 10569–78. https://doi.org/10.1523/JNEUROSCI.0091-11.2011.

Rudebeck, Peter H., and Elisabeth A. Murray. 2014. "The Orbitofrontal Oracle: Cortical Mechanisms for the Prediction and Evaluation of Specific Behavioral Outcomes." *Neuron* 84 (6): 1143–56. https://doi.org/10.1016/j.neuron.2014.10.049.

Russell, Paul. 2007. *Hume on Free Will*. December 14. https://plato.stanford.edu/archives/fall2021/entries/hume-freewill/.

Saddoris, Michael P., Fabio Cacciapaglia, R. Mark Wightman, and Regina M. Carelli. 2015. "Differential Dopamine Release Dynamics in the Nucleus Accumbens Core and Shell Reveal Complementary Signals for Error Prediction and Incentive Motivation." *The Journal of Neuroscience* 35 (33): 11572–82. https://doi.org/10.1523/JNEUROSCI.2344-15.2015.

Saunders, Blair, Hause Lin, Marina Milyavskaya, and Michael Inzlicht. 2017. "The Emotive Nature of Conflict Monitoring in the Medial Prefrontal Cortex." *International Journal of Psychophysiology*, The Psychophysiology of Motivation: Body and Brain in Action, vol. 119 (September): 31–40. https://doi.org/10.1016/j.ijpsycho.2017.01.004.

Scarantino, Andrea. 2016. "The Philosophy of Emotions and Its Impact on Affective Science." In *Handbook of Emotions 4th Edition*, 4th ed., edited by Lisa Feldman Barrett, Michael Lewis, and Jeannette M. Haviland-Jones. Guilford Publications.

Schoenbaum, Geoffrey, and Matthew Roesch. 2005. "Orbitofrontal Cortex, Associative Learning, and Expectancies." *Neuron* 47 (5): 633–36. https://doi.org/10.1016/j.neuron.2005.07.018.

Schoenbaum, Geoffrey, Matthew R. Roesch, Thomas A. Stalnaker, and Yuji K. Takahashi. 2009. "A New Perspective on the Role of the Orbitofrontal Cortex in Adaptive Behaviour." *Nature Reviews. Neuroscience* (London, United States) 10 (12): 885–92. https://doi.org/10.1038/nrn2753.

Schröder, Tobias, and Paul Thagard. 2015. "Emotions as Semantic Pointers: Constructive Neural Mechanisms." In *The Psychological Construction of Emotion*, edited by Lisa Feldman Barrett and James A. Russell. The Guilford Press.

Schroeder, Timothy. 2004. *Three Faces of Desire*. Philosophy of Mind Series. Oxford University Press.

Schultz, Wolfram. 2002. "Getting Formal with Dopamine and Reward." *Neuron* 36 (2): 241–63. https://doi.org/10.1016/S0896-6273(02)00967-4.

Schultz, Wolfram. 2006. "Behavioral Theories and the Neurophysiology of Reward." *Annual Review of Psychology* 57 (Volume 57, 2006): 87–115. https://doi.org/10.1146/annurev.psych.56.091103.070229.

Schultz, Wolfram, Peter Dayan, and P. Read Montague. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593–600.

Schultz, Wolfram, Léon Tremblay, and Jeffrey R. Hollerman. 2000. "Reward Processing in Primate Orbitofrontal Cortex and Basal Ganglia." *Cerebral Cortex* 10 (3): 272–83. https://doi.org/10.1093/cercor/10.3.272.

Shackman, Alexander J., Tim V. Salomons, Heleen A. Slagter, Andrew S. Fox, Jameel J. Winter, and Richard J. Davidson. 2011. "The Integration of Negative Affect, Pain and Cognitive Control in the Cingulate Cortex." *Nature Reviews. Neuroscience* (London, United States) 12 (3): 154–67. https://doi.org/10.1038/nrn2994.

Shadlen, Michael N., and William T. Newsome. 1996. "Motion Perception: Seeing and Deciding." *Proceedings of the National Academy of Sciences* 93 (2): 628–33. https://doi.org/10.1073/pnas.93.2.628.

Shadlen, Michael N., and William T. Newsome. 2001. "Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey." *Journal of Neurophysiology* 86 (4): 1916–36. https://doi.org/10.1152/jn.2001.86.4.1916.

Shenhav, Amitai, Matthew M. Botvinick, and Jonathan D. Cohen. 2013. “The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function.” *Neuron* 79 (2): 217–40.

Shenhav, Amitai, and Randy Buckner L. 2014. “Neural Correlates of Dueling Affective Reactions to Win-Win Choices.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (30): 10978–83.

Shine, James, M., and Russell A. Poldrack. 2018. “Principles of Dynamic Network Reconfiguration across Diverse Brain States.” *NeuroImage* 180: 396–405.

Shushruth, S., Mark Mazurek, and Michael N. Shadlen. 2018. “Comparison of Decision-Related Signals in Sensory and Motor Preparatory Responses of Neurons in Area LIP.” *The Journal of Neuroscience* 38 (28): 6350–65. https://doi.org/10.1523/JNEUROSCI.0668-18.2018.

Sinnott-Armstrong, Walter, and Lynn Nadel, eds. 2011. *Conscious Will and Responsibility: A Tribute to Benjamin Libet*. Series in Neuroscience, Law, and Philosophy. Oxford University Press.

Soon, Chun Siong, Marcel Brass, Hans-jochen Heinze, and John-dylan Haynes. 2008. “Unconscious Determinants of Free Decisions in the Human Brain.” *Nature Neuroscience* (New York, United States) 11 (5): 543–45. https://doi.org/10.1038/nn.2112.

Spivey, Michael. 2007. *The Continuity of Mind*. Oxford University Press.

Sporns, Olaf. 2011. “The Human Connectome: A Complex Network.” *Annals of the New York Academy of Sciences* 1224 (1): 109–25. https://doi.org/10.1111/j.1749-6632.2010.05888.x.

Strait, Caleb E., Tommy C. Blanchard, and Benjamin Y. Hayden. 2014. “Reward Value Comparison via Mutual Inhibition in Ventromedial Prefrontal Cortex.” *Neuron* 82 (6): 1357–66. https://doi.org/10.1016/j.neuron.2014.04.032.

Strauss, Gregory P., James A. Waltz, and James M. Gold. 2014. “A Review of Reward Processing and Motivational Impairment in Schizophrenia.” *Schizophrenia Bulletin* 40 (Suppl_2): S107–16. https://doi.org/10.1093/schbul/sbt197.

Sugrue, Leo P., Greg S. Corrado, and William T. Newsome. 2005. “Choosing the Greater of Two Goods: Neural Currencies for Valuation and Decision Making.” *Nature Reviews Neuroscience* 6 (5): 363–75. https://doi.org/10.1038/nrn1666.

Svoboda, Eva, Margaret C. McKinnon, and Brian Levine. 2006. “The Functional Neuroanatomy of Autobiographical Memory: A Meta-Analysis.” *Neuropsychologia* 44 (12): 2189–208. https://doi.org/10.1016/j.neuropsychologia.2006.05.023.

Tabibnia, Golnaz, John R. Monterosso, Kate Baicy, et al. 2011. “Different Forms of Self-Control Share a Neurocognitive Substrate.” *The Journal of Neuroscience* 31 (13): 4805–10. https://doi.org/10.1523/JNEUROSCI.2859-10.2011.

Takahashi, Yuji K, Matthew R Roesch, Robert C Wilson, et al. 2011. “Expectancy-Related Changes in Firing of Dopamine Neurons Depend on Orbitofrontal Cortex.” *Nature Neuroscience* 14 (12): 1590–97. https://doi.org/10.1038/nn.2957.

Tappolet, Tappolet. 2014. “Emotions, Reason, and Autonomy.” In *Autonomy, Oppression, and Gender*, edited by Andrea Veltman and Mark Piper. Studies in Feminist Philosophy. Oxford university press.

Thagard, Paul. 2006. “Desires Are Not Propositional Attitudes.” *Dialogue: Canadian Philosophical Review / Revue Canadienne de Philosophie* 45 (1): 151–56. https://doi.org/10.1017/S001221730000038X.

Thalberg, Irving. 1978. “Hierarchical Analyses of Unfree Action.” *Canadian Journal of Philosophy* (Edmonton, Alta., Canada) 8 (2): 211–26.

Thaler, Richard H., and H. M. Shefrin. 1981. “An Economic Theory of Self-Control.” *Journal of Political Economy* 89 (2): 392–406.

Thompson, Kirk G., Narcisse P. Bichot, and Jeffrey D. Schall. 1997. “Dissociation of Visual Discrimination From Saccade Programming in Macaque Frontal Eye Field.” *Journal of Neurophysiology* 77 (2): 1046–50. https://doi.org/10.1152/jn.1997.77.2.1046.

Thompson, Kirk G., Doug P. Hanes, Narcisse P. Bichot, and Jeffrey D. Schall. 1996. “Perceptual and Motor Processing Stages Identified in the Activity of Macaque Frontal Eye Field Neurons during Visual Search.” *Journal of Neurophysiology* 76 (6): 4040–55. https://doi.org/10.1152/jn.1996.76.6.4040.

Tremblay, Leon, and Wolfram Schultz. 1999. “Relative Reward Preference in Primate Orbitofrontal Cortex.” *Nature* (London, United States) 398 (6729): 704–8. https://doi.org/10.1038/19525.

Tsujimoto, Satoshi, Aldo Genovesio, and Steven P. Wise. 2011. “Comparison of Strategy Signals in the Dorsolateral and Orbital Prefrontal Cortex.” *The Journal of Neuroscience* 31 (12): 4583–92. https://doi.org/10.1523/JNEUROSCI.5816-10.2011.

Van Steenbergen, Henk. 2015. “Affective Modulation of Cognitive Control: A Biobehavioral Perspective.” In *Handbook of Biobehavioral Approaches to Self-Regulation*, edited by Guido H.E. Gendolla, Mattie Tops, and Sander L. Koole. Springer New York. https://doi.org/10.1007/978-1-4939-1236-0_7.

Velleman, James David. 2006. “Identification and Identity.” In *Self to Self: Selected Essays*. Cambridge University Press.

Vigneau, M., V. Beaucousin, P. Y. Hervé, et al. 2006. “Meta-Analyzing Left Hemisphere Language Areas: Phonology, Semantics, and Sentence Processing.” *NeuroImage* (Amsterdam, United Kingdom) 30 (4): 1414–32. https://doi.org/10.1016/j.neuroimage.2005.11.002.

Volkow, Nora D., Joanna S. Fowler, Gene-Jack Wang, et al. 2010. “Cognitive Control of Drug Craving Inhibits Brain Reward Regions in Cocaine Abusers.” *NeuroImage* (Amsterdam, United Kingdom) 49 (3): 2536–43. https://doi.org/10.1016/j.neuroimage.2009.10.088.

Volkow, Nora D., Michael Michaelides, and Ruben Baler. 2019. “The Neuroscience of Drug Reward and Addiction.” *Physiological Reviews* 99 (4): 2115–40. https://doi.org/10.1152/physrev.00014.2018.

Wager, Tor D., John Jonides, and Susan Reading. 2004. “Neuroimaging Studies of Shifting Attention: A Meta-Analysis.” *NeuroImage* (Amsterdam, United Kingdom) 22 (4): 1679–93. https://doi.org/10.1016/j.neuroimage.2004.03.052.

Wager, Tor D., and Edward E. Smith. 2003. “Neuroimaging Studies of Working Memory:” *Cognitive, Affective, & Behavioral Neuroscience* 3 (4): 255–74. https://doi.org/10.3758/CABN.3.4.255.

Walton, Mark E., Timothy E. J. Behrens, Mark J. Buckley, Peter H. Rudebeck, and Matthew F. S. Rushworth. 2010. “Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning.” *Neuron* 65 (6): 927–39. https://doi.org/10.1016/j.neuron.2010.02.027.

Watson, Gary. 1975. “Free Agency.” *The Journal of Philosophy* 72 (8): 205–20. https://doi.org/10.2307/2024703.
Watson, Gary. 2004. *Agency and Answerability: Selected Essays*. Clarendon Press ; Oxford University Press.
Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. MIT Press.
Westbrook, Andrew, Daria Kester, and Todd S. Braver. 2013. “What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference.” *PLoS ONE* 8 (7): e68210. https://doi.org/10.1371/journal.pone.0068210.
Winkielman, Piotr, Kent C. Berridge, and Julia L. Wilbarger. 2005. “Unconscious Affective Reactions to Masked Happy Versus Angry Faces Influence Consumption Behavior and Judgments of Value.” *Personality and Social Psychology Bulletin* 31 (1): 121–35. https://doi.org/10.1177/0146167204271309.
Winstanley, Catharine A., David E. H. Theobald, Rudolf N. Cardinal, and Trevor W. Robbins. 2004. “Contrasting Roles of Basolateral Amygdala and Orbitofrontal Cortex in Impulsive Choice.” *The Journal of Neuroscience* 24 (20): 4718–22. https://doi.org/10.1523/JNEUROSCI.5606-03.2004.
Wise, Roy A. 1980. “The Dopamine Synapse and the Notion of ‘Pleasure Centers’ in the Brain.” *Trends in Neurosciences* 3 (4): 91–95.
Wise, Roy A. 2004. “Dopamine, Learning and Motivation.” *Nature Reviews. Neuroscience* (London, United States) 5 (6): 483–94. https://doi.org/10.1038/nrn1406.
Wurtz, Robert H., and Michael E. Goldberg. 1972. “Activity of Superior Colliculus in Behaving Monkey. 3. Cells Discharging before Eye Movements.” *Journal of Neurophysiology* 35 (4): 575–86. https://doi.org/10.1152/jn.1972.35.4.575.
Yee, Debbie M, and Todd S Braver. 2018. “Interactions of Motivation and Cognitive Control.” *Current Opinion in Behavioral Sciences*, Emotion-cognition interactions, vol. 19 (February): 83–90. https://doi.org/10.1016/j.cobeha.2017.11.009.
Yin, Henry H., Barbara J. Knowlton, and Bernard W. Balleine. 2004. “Lesions of Dorsolateral Striatum Preserve Outcome Expectancy but Disrupt Habit Formation in Instrumental Learning.” *European Journal of Neuroscience* 19 (1): 181–89. https://doi.org/10.1111/j.1460-9568.2004.03095.x.

Glossary

Basal Ganglia (BG): A group of structures including: the ventral striatum (including the nucleus accumbens and olfactory tubercle), the dorsal striatum (including the caudate nucleus and putamen), the globus pallidus, the ventral pallidum, the substantia nigra, and the subthalamic nucleus. The BG has strong connections to both the cortex and the thalamus.

Cognition: Cognition refers to processes such as memory, attention, language, problem solving and planning (Pessoa 2008, 148).

Cortical Midline Structures (CMS): A group of structures including: ventromedial prefrontal cortex (vmPFC), perigenueal anterior cingulate cortex (PACC), dorsomedial prefrontal cortex (dmPFC), posterior cingulate cortex (PCC).

Executive Functions: Executive Functions are a set of cognitive processes that help support goal-directed behaviour. These processes support planning, attention, working memory, inhibition/control, and switching between tasks.

Emotion: At this point in time there is not a consensus on a neuroscientific definition for emotions. However, there are a few things that seem to capture what is meant by the term emotions. First, we can point out that emotions are internal states that cause certain changes such as, external changes in behaviour, internal physiological changes in the body, changes to our other mental states, and in some species (like humans), changes in our conscious awareness (Adolphs and Anderson 2018, 30). There are also certain properties of emotion states that are commonly discussed alongside emotions, such as: valence, arousal, scalability, persistence,

conscious feelings, and social communication (Adolphs and Anderson, 2018). Philosophically, the study of emotions can be said to fall under three general approaches: appraisal, feeling, and motivational theories (Scarantino 2016).

Motivation: An internal state that drives our goal-directed behaviours, either toward rewarding goals or away from aversive outcomes. While distinct from emotion, motivation is related to and can overlap with emotional properties. Motivation states can be triggered by either interoceptive or exteroceptive stimuli (Pessoa 2008; Adolphs and Anderson 2018).

Properties of Emotion States: Scalability, Valence, Persistence, generalization, global coordination, automaticity, social communication (Adolphs and Anderson 2018)

Subcortical: A term referring to the structures in the brain below the level of cortex, such as the amygdala, thalamus, hippocampus, and brainstem (Adolphs and Anderson, 2018).

Valence: One common property of emotions which describes the pleasantness or unpleasantness dimension (Adolphs and Anderson, 2018).