UNIVERSITY OF CALIFORNIA SAN DIEGO

The Mechanistic and Normative Structure of Agency

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Philosophy and Cognitive Science

by

Richard Jason Winning

Committee in charge:

Professor William Bechtel, Chair
Professor Nancy Cartwright
Professor Rick Grush
Professor Jeffrey Krichmar
Professor Piotr Winkielman

2019

The Dissertation of Richard Jason Winning is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

This dissertation is dedicated to my parents, Richard Lee Winning and JoLinda Winning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many people who played key roles in my development as a student and as a philosopher. One of the people who deserves the most credit is Mylan Engel. When I was an undergraduate at Northern Illinois University majoring in computer science and minoring in philosophy in 2002, I enrolled in Mylan's combined upper division/graduate survey on epistemology. It was during this course that began to really understand and appreciate academic philosophy as a discipline and as a skill. It was ultimately due to his encouragement and support that I enrolled in the graduate program in philosophy at NIU several years later. The time I spent as his student and as his teaching assistant have been the most influential experiences that have shaped the way I approach writing and teaching to this day.

Two other undergraduate professors in computer science who saw my potential more clearly than I did at the time are Kirk Yenerall and Robert Rannie. Yenerall helped to stimulate my interest in computer science as a theoretical subject and an academic discipline (not just a hobby or a way to earn a paycheck). He expressed confidence that I could excel as a teacher and that I was capable of earning a Ph.D. at a time in my life when such possibilities were almost inconceivable to me. Rannie pushed me harder than any other professor ever did and helped me to find the self-discipline, confidence, and ability to see past and overcome self-imposed limitations to reach my fullest potential as a student, as a programmer, and as a human being.

During my time as a master's student, the philosophy department at NIU provided the perfect environment to achieve the broad grounding of knowledge and skills in academic philosophy I would need to be able to hit the ground running in a Ph.D. program in two years, despite my lack of an undergraduate degree in philosophy. When I started there, I knew I was interested in philosophy of mind and philosophy of psychology. During these two years, my primary mentor was Carl Gillett; he was key to helping me to begin to understand the complex relationships between philosophy of mind, philosophy of science, and metaphysics. My abilities

as a writer and as a researcher expanded dramatically under his mentorship. He also showed me what it could look like to do high-quality philosophy in a way that does full justice to the importance and difficulty of problems in metaphysics while simultaneously engaging in a rigorous and direct way with empirical scientific research, and introduced me to the work of philosophers of science that would become central to my thinking, such as William Wimsatt and Bill Bechtel.

At UCSD, Rick Grush is undoubtably one of the people who were most important for my development. In particular he was a role model in getting me to see past the fashionable debates and research topics of the moment and explore more enduring theoretical ideas present in older writings that have not been fully appreciated. This advice especially paid off when I immersed myself into the writings of Howard Pattee, as well as the work of Michael Polányi, Donald MacKay, and theoretical work from the 1980s and 1990s on multi-agent artificial intelligence systems. These discoveries ultimately are what led to the development of this dissertation. Like Carl Gillett, Rick was also a role model as a philosopher who engages with science in rigorous and highly productive ways that lead to deep, novel insights about core philosophical issues.

My Ph.D. advisor, Bill Bechtel, also of course deserves a tremendous amount of credit. He has been a never-ending source of guidance, mentoring, advice, encouragement, and support. There have been countless times that he saw the potential value of one of my ideas before I did, and encouraged me and provided the necessary guidance to follow up on it. He is unique as an advisor in his willingness to help his students deeply explore even those subjects and ideas that he is not directly working on, and in his ability and willingness to provide continual critique and feedback that stimulate growth and maturity as an intellectual. He became my advisor at a time when I didn't envision philosophy of biology as a core research area of mine, but ironically it was I who steered myself into philosophy of biology after I became convinced of the unrealized potential of Pattee's ideas for addressing problems about agency

and other subjects. Bill's willingness to spend so much time exploring these possibilities with me and to read and discuss so many of Pattee's works ultimately made this dissertation possible and have also led to several collaborative publications. He is also an outstanding project manager; to the extent that I have been able to find the right balance between depth and breadth of scope in this dissertation (a seemingly impossible feat given its ambitions!), he deserves credit.

A partial list of some of the other people who have helped me to become a professional philosopher would certainly include Matt Babb, Matt Braich, Harold Brown, Dan Burnston, Nancy Cartwright, James Dye, Alicia Finch, Matt Fulkerson, Tanya Hall, Rebecca Hardesty, Tomis Kapitan, Anna Marmodoro, Noel Martin, Andrew Morgan, Doug Nitz, Geoff Pynn, Brock Rough, Gila Sher, Ben Sheredos, Eric Watkins, and Piotr Winkielman.

Chapter 2, in full, is a slightly expanded version of the material as it appears in "Mechanistic Causation and Constraints: Perspectival Parts and Powers, Non-Perspectival Modal Patterns," *British Journal for the Philosophy of Science*, forthcoming. The dissertation author was the sole author of this paper.

Chapter 3, in full, is currently under review for publication as a journal article under the title "Internal Perspectivalism: The Solution to Generality Problems about Proper Function and Natural Norms." The dissertation author was the sole author of this paper.

VITA

2003 Bachelor of Science, Northern Illinois University

2011 Master of Arts, Northern Illinois University

2019 Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

The Mechanistic and Normative Structure of Agency

by

Richard Jason Winning

Doctor of Philosophy in Philosophy and Cognitive Science

University of California San Diego, 2019

Professor William Bechtel, Chair

I develop an interdisciplinary framework for understanding the nature of agents and agency that is compatible with recent developments in the metaphysics of science and that also does justice to the mechanistic and normative characteristics of agents and agency as they are understood in moral philosophy, social psychology, neuroscience, robotics, and economics. The framework I develop is internal perspectivalist. That is to say, it counts agents as real in a perspective-dependent way, but not in a way that depends on an external perspective. Whether or not something counts as an agent depends on whether it is able to have a certain kind of perspective. My approach differs from many others by treating *possession of a perspective* as more basic than the possession of agency, representational content/vehicles, cognition, intentions, goals, concepts, or mental or psychological states; these latter capabilities require

the former, not the other way around. I explain what it means for a system to be able to have a perspective at all, beginning with simple cases in biology, and show how self-contained normative perspectives about proper function and control can emerge from mechanisms with relatively simple dynamics. I then describe how increasingly complex control architectures can become organized that allow for more complex perspectives that approach agency. Next, I provide my own account of the kind of perspective that is necessary for agency itself, the goal being to provide a reference against which other accounts can be compared. Finally, I introduce a crucial distinction that is necessary for understanding human agency: that between inclinational and committal agency, and venture a hypothesis about how the normative perspective underlying committal agency might be mechanistically realized.

Chapter 1
Introduction

The purpose of this dissertation is to develop a framework for understanding the nature of agents and agency that is compatible with recent developments in the metaphysics of science and that also does justice to the mechanistic and normative characteristics of agents and agency as they are understood in moral philosophy, social psychology, neuroscience, robotics, and economics. Here, 'agent' is understood as referring to an entity that can perform *actions* (as opposed to merely doing things); a human is an agent, whereas a toaster is not. Developing a cross-disciplinary framework for understanding agency is a difficult task because agency is not a univocal concept. In some contexts (e.g., where it may not even make much sense to speak of cognition, mentality, or consciousness), simple robots and organisms are counted as agents. In others, like economics and social psychology, agency is associated with complex capacities like rational deliberation and long-term planning. Rather than fixing on a single definition of agency, then, this dissertation develops a systematic framework of interrelated concepts in terms of which more or less demanding notions of agency can be defined and distinguished.

There are two assumptions that I make that crucially shape the way this dissertation proceeds. First, I assume that agents are *real*. Agents are not merely instrumentally useful fictions or posits. In fact, it would not make sense for this to be true: in order for there to be a posit, something (an agent) must be there to do the positing. So agents cannot themselves simply be posits. They must have objective existence that is independent of other minded beings.

The second assumption I make is that *agent* is not a natural kind. In other words, there is no single division between what counts as an agent and what doesn't that is the ultimately

"right" one, in a sense that is perspective-independent. Whether or not something counts as an agent is context-dependent. This is even true in a sense that goes beyond the mere observation that 'agent' is defined in different ways. On many of the ways 'agency' is defined in different fields, there will not be a perspective-independent way of drawing boundary lines around what counts as an agent and what doesn't.

The present work differs from other discussions of agency in that it takes both of these assumptions seriously and applies them consistently. Other accounts implicitly or explicitly relax one or both of these assumptions at various times. It might seem that there is no other choice but to do so. How can something be real in a mind-independent way, but also have an existence that is ineliminably perspectival? There is only one possibility: an agent must be capable of defining its own existence by means of its own perspective. And the possession of a perspective must be more basic than the possession of a mind: something can have a perspective without having a mind.

But how can something be real in a way that depends on itself being real? If something is grounded in itself, it might be objected, then it is grounded in nothing at all. The solution to this requires a key insight made by Daniel Dennett (1991): even when it is in question whether something is real *as an entity*, we can allow that it exists as a *real pattern*. The key, then, is that a *perspective* can emerge merely by means of the emergence of a certain kind of *pattern*.[1] Once this happens, the perspective can ground the existence of other things. The reality of an agent *as an entity* depends on its reality *as a pattern*. Chapters 2 and 3 provide substantiation and argumentation for this line of reasoning.

Among other things, an agent is a kind of mechanism. A person is a mechanism, in the sense that a person's behavior is explainable in terms of the causal organization of his or her parts. The geyser Old Faithful is also a mechanism in this sense (though in the case of Old

[1] On the difference between the emergence of *being* (or the emergence of an *entity*) and the emergence of a *pattern*, see Winning and Bechtel (forthcoming).

Faithful we don't refer to the causal organization of its parts as its "physiology"). It makes sense to ask for the reason why Old Faithful erupts on a given occasion, just as it makes sense to ask for the reason why a person stands up on a given occasion. But whereas it also makes sense to inquire into the *person's reason* for standing up, it would not make sense to inquire into *Old Faithful's reason* for erupting. This is because whereas persons are agents, Old Faithful is not an agent. Whereas Old Faithful engages in behaviors (i.e., erupting), Old Faithful does not perform *actions*; only agents do that. What is distinctive about agents is that some of their behaviors count as actions, i.e., behaviors that are performed for the agent's own reason.[2]

We often treat many kinds of organisms as agents (i.e., as if they have their own reasons for doing what they do). Certain kinds of artificial intelligence systems and robots might also be viewed as having reasons for what they do (especially those that are designed to engage in means-end reasoning or deliberation). But in general, machines and inanimate objects don't have their own reasons for their behaviors. A toaster doesn't toast bread *for* any reason at all (though its human user may have a reason in mind); it just does it.

One might be tempted to say that the toaster can't have its own reason for doing anything because its causal organization fully constrains its behavior. But this is also true of persons: human beings are of course fully constrained by their causal organization (i.e., their physiology) to behave the way they do as well. In fact, all agents are mechanisms: all organisms and artificial systems that count as agents are similarly constrained to act the way they do by the causal organization of their parts. What, then, is the special nature of the causal organization of certain mechanisms in virtue of which they count as agents, i.e., what kind of *pattern* must emerge for there to exist an agent? That is the question with which much of this

[2] When speaking about the agent's own reasons, the reasons I am referring to are agent-relative (as opposed to agent-neutral; Portmore, 2013) and motivating (as opposed to moral, or what Parfit, 1997 calls "normative") reasons.

dissertation is concerned. In the rest of this introduction, I provide an overview of how the chapters unfold.

Chapter 2 draws from the work of New Mechanist philosophers of science to describe reality as composed of an ontologically dense "rainforest" (as Ross, 2000 puts it) of modal real patterns that are not, as Craver (2013, p. 140) puts it, "prechunked into mechanisms" (nor, for that matter, into objects and properties). What counts as a mechanism often depends on the context of explanation and on the explanatory interests of the biologist in question. Mechanisms therefore only exist, or are "real," *as mechanisms*, only in virtue of having been identified as such, usually by biologists (though they are real *as patterns* in an unqualified sense). Chapter 2 explores the metaphysical ramifications of this and presents an ontological framework, grounded not in philosophical armchair speculation but in scientific practice, centered on the ontological category of *constraints*, a notion that was foundational to the development of later forms of classical mechanics as well as quantum mechanics, but mostly neglected by philosophers. The existence of constraints is not dependent on perspectives. But when we define boundaries around systems of constraints and consider them at certain levels of detail, perspectivally-dependent causal powers, objects, properties, and mechanisms emerge. Chapter 2 spells out how the characteristics possessed by mechanisms and their parts (including those that allow for mechanistic explanation) can emerge from the ontological bedrock of constraints combined with the perspective-taking of external observers.

As noted earlier, this would seem to present a conundrum: if agents are mechanisms, and mechanisms are ontologically dependent on perspectives, but perspectives depend on agents (i.e., somebody has to do the perspective-taking), then agents seem locked in an ontological chicken-and-egg paradox. If agent A only exists because of agent B's perspective, and agent B only exists because of agent C's perspective, etc. then how does the entire chain of agents exist? Is there a perspective-taker whose existence is not grounded in a perspective? I argue that this regress problem can only be solved by dropping a key assumption: that of

external perspectivalism. External perspectivalism says that if the existence of X is dependent on Y's having a certain perspective, then Y must be (physically or ontologically) external to X. The denial of this thesis yields *internal perspectivalism*: X can be what has the perspective that X's own existence as an instance of a perspective-dependent ontological category C (e.g., mechanism) depends on. In other words, X's possession of a certain perspective can ground X's own ontological characteristics. Something like this idea is central to the Autopoietic approach to theoretical biology (e.g., Varela, 1997).

But now, internal perspectivalism might seem to be describing things as ontologically grounded in nothing: the perspective-haver and the perspective-having are mutually grounding, but their mutual grounding flows in an infinite cycle that never seems to be grounded by anything else. The key here is to realize that the perspective-haver, like everything, also has a non-perspectival existence *as a real pattern*, i.e., as a particular organization of *constraint.* What is distinctive about those systems of constraint that define their own perspectival existence must be that they possess a certain distinctive kind of organization. My approach differs from many others by treating *possession of a perspective* as more basic than the possession of agency, representational content/vehicles, cognition, intentions, goals, concepts, or mental or psychological states; these latter capabilities require the former, not the other way around. But a crucial question emerges: what is the minimal organization that is necessary and sufficient for a system to have its own perspective?

To begin to understand the minimal conditions for a physical system to have a perspective, I turn in Chapter 3 to the philosophical debate about the nature of biological proper functions. The practice of biologists seems to presuppose objective divisions between what counts as proper versus improper function for biological systems. This is in fact key to the mechanistic perspectives they take that are discussed in Chapter 2. I examine a number of non-perspectival attempts to analyze proper function, finding that they fail because it is impossible to analyze a normative standard in a way that does not make reference to a

perspective. As I explain in Chapter 3, a perspective is needed to establish the categories of entities to which the normative standard applies; a normative standard can only apply to a world that has been ontologically "carved" (or in Craver's language, "chunked"), i.e., it can only apply to a scheme that establishes which real patterns will be counted as "the same" and which will not.

But I also argue that the activity of "natural selection" does not imply any fixed perspective, and attempts to ground biological proper function in the external perspectives of scientists also fail. Proper function within a biological system ultimately must be grounded in that system's own ability to have a perspective on what it is the job of some part(s) or trait(s) to perform, and how to intervene when they are not performing their job properly. I then describe how this is possible: certain types of biological molecules are capable of sorting other objects in their surroundings into discrete types, and enacting categorical responses depending on which type the objects they come into contact with fall into. These molecules are organized into larger schemes that test for certain conditions and enact differential responses depending on which condition is present (i.e., to leave things alone if they are functioning normally, or intervene if they are not). I explain the sense in which such organized activities can provide an internal perspectivalist, determinate grounding for a normative standard of functioning within biological systems.

In Chapter 4, I attempt to draw a more general lesson from the special case of proper function normativity, to argue that any system that grounds a normative standard by means of its own internal perspective must possess a certain type of organization: in my terminology, it must count as an *observer-worker system*. I explain the sense in which an observer-worker system can ground the most basic kind of normative standard that can exist physically: a standard of what counts as *thermodynamical work*. Such a system also counts as an *observer* in a way that is independent of whether an external observer considers it as such. An observer-worker system minimally classifies events in its environment according to a discrete

categorization scheme, and produces behavioral responses that are only sensitive to such events at the level of detail of their category type (i.e., its behavior is not sensitive to difference between events past a certain level of detail). As I explain, this way of understanding the fundamental requirements for the grounding of a normative system owes a great deal to the work of Howard Pattee.

I then go on in Chapter 4 to define another type of normativity that (like proper function normativity) is a special case of observer-worker normativity, but that is distinct from proper function normativity. This is the normativity of *control*. To say that X controls Y means more than simply that X causally influences Y: it means that X's causal influence exceeds a certain type of threshold (sufficient to separate it from being "mere influence" rather than control). This way of defining control is inspired by the work of Nicholas Rescher. An *autonomous controller* is one that has its own perspective on where this threshold lies. A large part of Chapter 4 is then devoted to describing a number of kinds of organization that a system can possess in virtue of which it can ground such normative standards.

Agents are a certain kind of autonomous control system, but they are a very sophisticated kind.[3] In fact, there are a number of types of autonomous controllers that are referred to as 'agents' or that are ascribed characteristics that are associated with agency, such as goal-directedness. Workers in different disciplines do not universally agree on what kinds of control systems should be called 'agents', and this fact sometimes makes it difficult for ideas to be translated from one discipline to another. In Chapter 4, I focus on the kinds of features that control systems can have in virtue of which they can possess control perspectives that increasingly approach the kind of sophistication that human agents possess. The goal here

[3] It should be noted here that 'autonomous' is used with respect to agency in many contexts with a different meaning than the one used here. For example, "autonomous" robots are those that can set their own top-level goals or that can function independently in a given environment. Sometimes 'autonomous' refers to a person's ability to make their own decisions. My usage of 'autonomous' is much more basic, and simply refers to any control system that has its own internal perspective on what kind of influence counts as control.

is to create a more neutral vocabulary so that various ways of understanding agential concepts from different domains can be compared and translated. As Davidson writes,

> We have many vocabularies for describing nature when we regard it as mindless, and we have a mentalistic vocabulary for describing thought and intentional action; what we lack is a way of describing what is in between. This is particularly evident when we speak of the "intentions" and "desires" of simple animals; we have no better way to explain what they do. (1999, p. 11)

Chapter 4 attempts to fill this lacuna.

If there is one characteristic that is most common across the diverse ways that agency is understood, it is that agents are fundamentally systems capable of making *choices*. Chapter 5 is based on the basic presupposition that the ability to make *choices* requires the ability to possess a distinct kind of perspective (beyond those considered in Chapter 4), and therefore a distinct kind of organization. All control systems make *selections*, but this is not the same thing as making a *choice*. In particular, when making a choice, what gets selected is a *state of affairs* against other states of affairs, on the basis that the chosen state of affairs is *preferred* over the others.

The key to understanding the capacity for choice is then to understand what it means to have a perspective that includes states of affairs, and on which some states of affairs are preferred over others. What I argue in Chapter 5 is that this involves having a perspective on which some things are potentially under my control sometimes and not under my control at other times. In this way, a representational content is not merely decoupled from particular behaviors; it is semantically decoupled from *control* as such. This represents the point where a system can have a perspective on something that does not essentially connect it to whether or how it is under the system's control. I argue that this is the emergence of *objective* representational contents. By representing this way, a system is able to have not merely its own perspective, but its own *reasons*. It can have a preference of one state of affairs over another, where the preference itself does not imply what must be done to obtain that state of affairs. Other objective representations of causal relations between states of affairs must be

combined with representations of potential actions and their causal consequences to engage in *practical reasoning* to obtain the preferred state of affairs.

This results in a relatively demanding definition of agency that, when combined with the broader taxonomy of agent-like control system architectures of Chapter 4, can provide a rich vocabulary for understanding how different approaches to understanding agency are related. However, they leave out a crucial phenomenon that is mostly associated with human agency: *commitment*. Humans are capable not merely of objective representations and acting on reasons pertaining to the situation at hand, but they can also form and act on commitments (e.g., plans or promises) that constrain how they will reason and act in the future. In Chapter 6, I argue that this is a fundamentally distinct type of perspective that cannot be explained by means of the machinery introduced in Chapters 4 and 5. The difference here is that the agent's own self, *as agent*, becomes not just the subject but the object *of its own agency*: it becomes one of the things that has been, and will be, either under its control or not under its control in various ways at various times. It doesn't just see its own body or its cognitive capacities as tools for its own use: it sees its own *agency* as a tool for its own use. And it can subject itself to its own past commands, as well as subordinate its future self to its current self. In Chapter 6, I not only explore what this means conceptually but also forward a hypothesis about the nature of how a system could be mechanistically organized to realize such a perspective.

There are a number of core philosophical debates that, although relevant to the issues at hand, I will not have space to directly address in this dissertation. The debate at the center of the branch of philosophy known as the philosophy of action is the "problem of deviant causal chains," i.e., the problem of defining when a movement or a behavior counts as an action. The debate at the center of the branch of philosophy known as epistemology is the problem of defining when a true belief counts as knowledge. The debate at the center of the branch of philosophy known as the philosophy of mind is the "mind-body problem," i.e., the problem of defining the nature of minds and how they are related to non-mental entities. Three of the core

problems within moral philosophy are the problems of personal identity and of how freedom of the will and moral responsibility are compatible with the universe as existing in the causal framework depicted by science. However, this dissertation does address the basic puzzle that underlies all of these problems and that has made their solution so elusive over the centuries: the puzzle of understanding how something can simultaneously be mind-independently real, ineliminably dependent on a perspective, and capable of providing a non-normative grounding for normative properties. What is needed to solve this puzzle in any given case is an *observer-worker system* (introduced in Chapter 4). The present framework, then, not only can facilitate interdisciplinary work on agency but can also pave the way for substantial progress to be made on these core debates in future work.

Chapter 2
Mechanistic Causation and Constraints: Perspectival Parts and Powers, Non-Perspectival Modal Patterns

2.1 Introduction

Mechanistic explanation is ubiquitous in science, and philosophy of science has been making great progress in gaining a realistic understanding of the epistemic practices of scientists. Less progress has been made in understanding the metaphysical implications of these new insights about mechanistic explanation. I argue that this is due to the fact that as a result of the insights gained about mechanistic explanation, there are now at least five key desiderata that must be satisfied by any account of its metaphysical underpinnings, and no extant account has managed to satisfy all five. In this chapter, I lay out these five desiderata and explain why existing accounts of the metaphysics of mechanistic causation fail to satisfy them. I then present an alternative account which does satisfy the five desiderata. According to this alternative account, we must resort to a type of ontological entity that is new to metaphysics, but that has been familiar to scientists for two centuries: constraints. In this chapter, I explain how a constraints-based metaphysics fits best with the emerging consensus on the nature of mechanistic explanation.[4]

[4] I do not discuss it in this chapter, but there is arguably a sixth desideratum: the need to account for control relationships in biological mechanisms. To see how the present constraint-based framework rises to this challenge, see Winning and Bechtel (2018).

2.2 Five Desiderata for an Account of Mechanistic Causation

The standard scientific realist ontology has it that objects (here conceived broadly enough to include entities such as particles and fields) exist, they have properties, and the universe has laws that determine what happens to these objects and properties over time. Scientists discover these laws and are able to explain observations of objects and properties and how they change (that is, phenomena) by reference to the laws.

Much has been written recently about the fact that, *prima facie* at least, this picture does not fit the epistemic practices of scientists in a number of domains, for example in biology. Biologists do not generally seek out or refer to laws to explain phenomena (Smart, 1963, pp. 50–61; Bechtel & Abrahamsen, 2005, p. 422); instead, biological explanations often appeal to mechanisms (Polányi, 1958, p. 357; Wimsatt, 1976, p. 671). What are the implications of this fact for the scientific realist ontology? One might respond to this fact by denying scientific realism in general, or denying realism about mechanisms in particular.[5] Perhaps scientific explanations in general, or mechanistic explanations in particular, are not true by virtue of referring to real entities in the first place. Maybe biologists' mechanisms are mere metaphors or useful fictions.

No attempt will be made in this chapter to defend any stance on debates between realism and antirealism. Instead, I want to begin by adopting the most straightforward metaphysical interpretation of the practice of biologists and granting the New Mechanist view that "mechanisms are real systems in nature" (Bechtel, 2006, p. 33), in order to see what type of view about causation squares best with treating mechanisms as real in some robust sense.

[5] The question of whether or not mechanisms are real is orthogonal to the debate about the metaphysics of explanations, i.e., whether explanations themselves consist of causal structures (the "ontic view" of explanations) or representations (the "epistemic view"; see Glennan, 2017, Section 8.2 for an excellent overview of this debate).

What is a scientific realist to conclude from the fact that biology explains in terms of real mechanisms rather than laws? When we look at the reasons why New Mechanists have rejected conventional accounts of causation,[6] the five desiderata will emerge.

2.2.1 The First Two Desiderata: Intrinsicness and Productivity

One of the themes of the New Mechanist literature is that mechanisms and their components are productive of changes, and that this productive nature is intrinsic to them. A mechanistic explanation does not merely describe occurrent regularities in the way a system changes over time and posit some type of external principle of change, such as laws, for the purpose of assigning responsibility for the production of these changes. Instead, with mechanistic explanation, what is responsible for producing changes lies within the mechanism itself, and the buck stops there. Machamer, Darden, and Craver (henceforth MDC) explain that in this sense, mechanistic explanation implies the rejection of a certain type of Humeanism about causation:

> We should not be tempted to follow Hume and later logical empiricists into thinking that the intelligibility of activities (or mechanisms) is reducible to their regularity. … Rather, explanation involves revealing the *productive* relation. (2000, pp. 21–22)

MDC argue that traditional ontological categories such as that of object, property, or process, which allow us to talk about the intrinsic nature of mechanisms, are not sufficient for characterizing the productive nature of mechanisms, and they argue that we must make room for activities as a category in our ontology:

[6] New Mechanists like Craver (2007) have sometimes adopted Woodward's (2003) manipulability account of causal explanation, but it is important to keep in mind that Woodward's account is intended only to be an account of "how we think about, learn about, and reason with various causal notions and about their role in causal explanation" (2008, p. 194), not an account of the metaphysics of causation. Additionally, Glennan has argued that the virtues of Woodward's account of the epistemology of causal reasoning "[do] not legitimate the manipulability theory as a metaphysical account of causation" (2009, p. 318).

> … it is artificial and impoverished to describe mechanisms solely in terms of entities, properties, interactions, inputs-outputs, and state changes over time. Mechanisms do things. They are active and so ought to be described in terms of the activities of their entities, not merely in terms of changes in their properties. (2000, p. 5)

MDC argue that activities themselves underwrite the modal characteristics (for example, counterfactuals) of causal relations, being "the producers of change" (2000, p. 4), and that therefore, "[n]o philosophical work is done by positing some further thing, a law, that underwrites the productivity of activities" (2000, p. 8).

But there is a further ontological alternative to activities that MDC must address: causal powers (also known as 'capacities' or 'dispositions'). MDC give short shrift to this alternative conception in their 2000 paper. Their remarks are confined to the following:

> Substantivalists thus speak of entities with capacities (Cartwright, 1989) or dispositions to act. However, in order to identify a capacity of an entity, one must first identify the activities in which that entity engages. One does not know that aspirin has the capacity to relieve a headache unless one knows that aspirin produces headache relief. (2000, pp. 4–5)

MDC therefore conclude that it is activities that are fundamental, and dispositions or capacities are ontologically secondary to them. Here, MDC have made the mistake of conflating epistemic priority with metaphysical priority. Cartwright has argued that knowledge of capacities does not reduce to knowledge of activities: "The knowledge we have of the capacity of a feature is not knowledge of what things with that feature do but rather knowledge of the nature of the feature" (1999, p. 78). But even if it were necessary to identify the occurrence of an activity before one could identity the capacity to engage in the activity, as MDC argue, it would not follow that the capacities are ontologically dependent on activities, and activities are therefore more fundamental.

Machamer makes a slightly different argument in a later paper:

> activities are better off ontologically than some people's ontic commitments to capacities, dispositions, tendencies, propensities, powers, or endeavours. All these concepts are derivative from activities. … [T]he active exercise of a capacity has to be ontologically prior to any mysterious property called "the ability to exercise that capacity." (2004, p. 30)

Instead of arguing from epistemic priority to ontic priority, Machamer is arguing from conceptual priority to ontic priority. Machamer explicitly states the underlying premise of this argument on the same page: “being able to recognize what a capacity does when actualized or the activity that constitutes it presupposes having the concept of the activity” (2004, p. 30). Again, there is a conflation at work: conceptual priority is not the same thing as ontological priority. But further, it’s not clear that the premise itself is true. As Aristotle argued, it is only the conception of a capacity (or “potentiality”) that allows us to make sense of an entity persisting through a change in its activity (Gill, 1989, pp. 185ff). Hence, it’s not obvious that activities are conceptually prior to capacities.

The above arguments are not the primary reason why powers and capacities have been mostly avoided by philosophers of science. The primary reason was alluded to in the quotation from Machamer: causal powers and their ilk are often viewed by philosophers of science as something to be avoided, because they are viewed as mysterious as well as alien to science. This brings us to the next desideratum.

2.2.2 The Third Desideratum: Scientific Validity/Non-Mysteriousness

While Cartwright and Pemberton (2013) have argued in favor of a robust, Aristotelian account of causal powers to ground the inherent activeness of mechanisms, and others[7] have understood biological causation in terms of causal powers, the New Mechanist philosophers have generally been unwilling to fully embrace an Aristotelian picture of causation that imbues the entities within mechanisms with causal powers.[8] There are a couple of reasons for this.

[7] For example, Gillett (2007), Dupré (2007), Deacon (2011, pp. 364–368), Mumford and Anjum (2011, pp. 218–220), and Moreno and Mossio (2015).

[8] Glennan at times writes of capacities and powers (e.g., 2017, pp. 31–35) but they are not basic categories in his ontology and take a backseat to activities (e.g., 2017, pp. 50 & 148). This is apparently because Glennan denies that capacities are ‘intrinsic features’ of their bearers (2017, p. 52). But on the

First, scientists don't tend to talk in a way that makes it obvious that they are committed to the reality of entities like powers.[9] Second, it has sometimes been argued (for example, by Machamer, 2004) that causal powers are mysterious. Why believe that causal powers are more mysterious than activities (that is, the actual behavior)? Because causal powers are dispositional. This fact, by itself, has often led philosophers to claim that causal powers are ontologically secondary to occurrent events and properties.

But what is it about dispositions that makes them mysterious? Consider what Nelson Goodman had to say on this subject:

> The peculiarity of dispositional predicates is that they seem to be applied to things in virtue of possible rather than actual occurrences—and possible occurrences are for us no more admissible as unexplained elements than are occult capacities. The problem, then, is to explain how dispositional predicates can be assigned to things solely on the basis of actual occurrences and yet in due accordance with ordinary or scientific usage. (1954, p. 42)

On Goodman's understanding, then, that which is merely dispositional cannot be actual; there is no logical space for any alternative. But Heil (2005, 2012) argues that this is not the only way to understand the meaning of 'dispositional'.[10] When powers are conceived of as dispositions, he argues, they are conceived as something that is actual, part of the intrinsic nature of its bearers, even when not manifested. Similarly, for Harré and Madden, having a power versus not having power is not merely a difference in what objects can do, or might do, and it might not be a difference in what they will do at all. It is rather "a difference in what they themselves now are … a difference in intrinsic nature" (1975, p. 86).

Heil's way of understanding 'dispositional' opens up a way out of Machamer's mysteriousness objection. But what could it mean for something to be part of the actual,

---

other hand, and interestingly, Glennan does characterize modal "relations of causal determination" as "intrinsic actual-world" features (2017, pp. 167–168).

[9] Cartwright has argued that a commitment to capacities or powers is "implicit … in the conventional methods for causal inference" (1989, p. 142) used by scientists, but many philosophers of science have remained unconvinced.

[10] See also Weissman's (1965, pp. 84–85) criticism of Goodman's view, which anticipates Heil's argument in important respects.

intrinsic nature of an object, but in a non-manifested state? And can this be understood in a way that is grounded in the details of scientific practice? A positive account that answers these questions will have to wait until Section 2.4. I turn now to the remaining two desiderata.

2.2.3 The Fourth Desideratum: Directionality

Mechanistic activities are sometimes characterized as being directional. Among the New Mechanist philosophers of science, Bechtel has laid the greatest stress on the directional aspect of the activeness of mechanisms:

> The term *activity* … does not readily capture the fact that in most operations there is also something acted upon. This is the reason I have preferred the term *operation*. Typical of the operations I have in mind are the reactions of chemistry which prototypically involve a catalyst, a reactant, a product, and often a cofactor. (2006, p. 30; see also Bechtel, 2008, p. 14)

In a similar vein, Glennan writes that "with respect to activities that are interactions, we often distinguish between the active 'doer' of the activity and the passive object of the activity" (2017, p. 31).[11] In other words, one component might play a passive role in a mechanistic operation, while another plays an active role. For example, a chaperone molecule in a cell performs the operation of folding a protein. After folding is completed, the bonding structure of the protein has changed, but the chaperone returns to its original configuration after the operation is complete. The chaperone plays an active role in the process: it changes the structure of the folding substrate without itself being substantially altered in the process (any alterations to the chaperone are readily reversed so that it can play the same role on subsequent occasions).

Several accounts of causation provide a natural interpretation for the directionality of causation. On the conserved quantity theory (Dowe, 2000), for example, one might explain

[11] Glennan further adds the caveat that "this distinction is not a deep one, in the sense that all actors in an interaction produce changes in other actors, and often the active/passive distinction is a matter of degree" (2017, p. 31). In this respect, the directionality of activities manifests one of the ways that mechanisms are perspectival, as discussed in Section 2.2.4.

directionality by reference to the fact that one component in an operation gains a quantity of energy while another loses a quantity of energy. While this might be useful for identifying the components playing an energetically active versus passive role in metabolic processes in biology (the molecule supplying the energy for a reaction might be seen as the active component), it would fail with respect to many signaling or regulatory types of processes. Any account of mechanistic causation should be able to provide a substantive story about the directionality of causation that is consistent with scientific practice.

This could potentially be another respect in which the causal powers approach has a leg up. A distinction between active and passive causal powers goes at least back to Locke, and has been revived in the form of 'powers' versus 'liabilities' by Harré and Madden (1975, p. 89), as well as 'backward-looking' versus 'forward-looking' powers by Shoemaker (1998).[12]

2.2.4 The Fifth Desideratum: Perspectival Nature of Mechanisms

Another one of the recurring themes in the New Mechanist literature is that mechanisms are perspectival: the boundaries and identity conditions of mechanisms are fixed, in part, by the subjective mental states (for example, interests or perspectives) of scientists.[13] Specifically, what counts as a mechanism is determined not by "what the mechanism invariably does but what we think it is supposed to do" (Craver, 2013, p. 140). Craver and Bechtel write that "there are no mechanisms simpliciter—only mechanisms for phenomena. A mechanism's phenomenon partially determines the mechanism's boundaries (i.e., what is 'in the mechanism and what is not)" (2006, p. 469), but as Darden points out, "the choice of phenomenon is

[12] It should be noted that a distinct notion of "directionality" or "directedness" pertaining to dispositions was discussed by Martin and Pfeifer (1986) and Molnar (2003). 'Directionality' in their sense merely refers to the relation between a disposition and its manifestations.
[13] This point was made earlier by Polányi (1958, p. 357), Kauffman (1971, pp. 259–260), and Glennan (1996, p. 52).

relative to the scientist's interests" (2008, p. 960). Craver emphasizes that "the world does not come prechunked into mechanisms"; such chunking ineliminably will be contextually dependent on the phenomenon of interest, that is, "some behavior that the scientist is interested, for whatever reason, in explaining" (2013, p. 140). As a result, Craver characterizes his view as a kind of "perspectivalism about mechanisms":

> Mechanistic and functional descriptions … presuppose a vantage point on the causal structure of the world, a stance taken by intentional creatures when they single out certain preferred behaviors as worthy of explanation. (2013, p. 134)

Kuhlmann and Glennan write that "the New Mechanists agree that there is an inherent perspectivalism in the process of identifying and individuating parts" (2014, p. 339), which in turn is key to the individuation and identity conditions of the mechanisms themselves.

This fifth desideratum, which has perhaps been most neglected in extant discussions of mechanistic causation, creates special difficulties for any attempt to provide a metaphysical account of mechanisms. Mechanistic processes are causal processes, and mechanisms are individuated relative to scientists' explanatory interests. But surely, causation itself is objective and mind-independent (otherwise, an enormous amount of ink has been spilled on the mind-body problem and the causal exclusion problem for nothing!). Further: above, I said that I would take for granted the New Mechanist view that "mechanisms are real systems in nature." How do we square this with the "inherent perspectivalism" of the individuation and identity conditions of mechanisms?

In the discussion leading up to Section 2.2.4, I have been drumming up the merits of causal powers as an approach to understanding the metaphysical underpinnings of mechanisms. But the causal powers approach does not offer much help once we consider the perspectival nature of mechanisms, and if anything, the causal powers approach only compounds the difficulties. MDC characterize causal powers approaches as fundamentally substantivalist; they are correct in the sense that an attribution of causal powers typically presupposes some scheme of individuation. On the usual picture of causal powers, they are

instantiated by individuals. To the extent that the individuation of mechanistic parts is dependent on the perspectives taken by scientists, then, the individuation of causal powers will be as well. In other words, powers are relative to a stance that one takes about what counts as an individual and what does not.[14]

Among the traditional options, causal powers are the best way to conceptualize mechanistic causation, but they cannot metaphysically be the end of the story. Instead, a new approach is needed that provides the necessary metaphysical foundation for understanding both causal powers and mechanistic causation and also satisfies the five desiderata. It is to this new approach that I now turn.

2.3 Constraints and Causation

2.3.1 Terminology

2.3.1.1 Definition of 'pattern' or 'structure'

For what follows, it is important first to get clear on the notion of a *pattern* or *structure* (I will treat these two terms as synonymous). A pattern is a repeatable way that things, parts, or portions of some kind of stuff (where 'stuff' is intended very broadly to include matter, energy, space, time, etc.) can be arranged or related in some way (when the underlying stuff is discretely divisible it is often referred to as the pattern's "elements"). Something can exist as a pattern even if it does not satisfy the requirements of any other ontological category (examples of such ontological categories are object, property, event, etc.). All of the socks that I have ever lost, considered collectively, instantiate a pattern. They each exist somewhere (let's assume

[14] Additionally, Heil (2012, pp. 118–119) argues that distinctions that are made between active and passive roles for causal powers in causation are ineliminably perspectival even if powers themselves are not.

they are still intact) and in spatial relations with one another. Another set of socks could, in principle, also become spatially arranged in that way. But though there is a pattern in the set of socks I have lost—a *real* pattern—there is no real *object* whose parts those socks are.

An important distinction about patterns is that between pattern *types* and pattern *tokens*. Only pattern types are repeatable; a pattern token is a single instance of the type and can only exist once. At any moment, for a pattern token to exist, there must be some kind of entities, material, or underlying stuff that the pattern is a pattern "of" at that moment. But at least for some kinds of patterns, the underlying stuff that instances a pattern can change over time while the same pattern token persists (as in the famous Ship of Theseus example).

Patterns exist at multiple levels of grain. Suppose an ordinary person draws a circle on a sheet of paper. We could build a detection device that detects circles at a very coarse level of grain, and another one that makes much finer discriminations. The first device might detect a circle on the sheet of paper, whereas the "imperfections" or "noise" in the circle (that are due to the imperfections of the medium and of the artist) may prevent the second device from counting it as a circle. They are in fact sensitive to two different patterns: the first admits of a greater range of variation in its instances than the second. Similarly, the set of socks I have lost do not merely instantiate one spatial pattern, they instantiate a range of patterns of different levels of coarseness of grain.

Pattern types may or may not include the nature of their elements or underlying stuff as part of their identity conditions. A sock-pattern (such as the one cited above) can only be instantiated by another set of socks. If a set of shoes became spatially arranged in the same way, they would instance the same spatial pattern but not the same sock-pattern (they cannot instance *any* sock-pattern; they are not socks). Consider two philosophers: A, who believes that the only things that are real are atoms and subatomic particles, and B, who believes that macro-scale objects are real. B points to a set of rocks arranged into a circle. A agrees that it is a circle, and that they are pointing to the same pattern. But A disagrees that it is a pattern of

rocks; A believes that it is a pattern of atoms, and not rocks. The pattern is the same, even though A and B disagree on how to ontologically categorize the medium that instantiates it. In fact, they might even agree that the pattern A sees is at the same level of grain as the one B sees. It is this very observation that has given rise to the "structural realist" approach in philosophy of science, one version of which says that scientific progress mostly consists in progress in terms of learning structural facts, whereas science has a poorer track record when it comes to the discovery of facts about fundamental ontological categories.

2.3.1.2 Definition of 'causal pattern' or 'causal structure'

Also important for present purposes will be the notion of a *causal pattern* or *causal structure* (I also treat these as synonymous). In a causal system, some patterns (e.g., spatial patterns of a certain type of material) cannot help but give rise to certain other kinds of patterns at some time in the future. The result is a causal pattern: a repeatable way that patterns within the causal system will (by some sort of necessity or modal force weaker than logical necessity) change over time. In a Newtonian causal system containing only two masses at rest, for example, the two masses will attract one another. Their spatial arrangement at time t1 is constrained to change into a certain different arrangement at time t2: the two masses will become closer to one another. This causal pattern will be repeated with any other system of two masses at rest.

Causal patterns, like any pattern, can exist at different levels of grain, and may or may not depend on their elements having certain properties. Causal patterns that are relations of patterns of a relatively coarse level of grain (like patterns of temperature or pressure) are more likely to be described in terms of probabilities instead of strict rules.

If the causal pattern itself is thought to include a modal aspect (which the "Humean" would deny), then the pattern, to be a causal pattern, will be partly constituted by whatever gives it that modal force—e.g., powers or laws. If we abstract away from the power or law, it

becomes only a pattern of what *did* happen in the system, not a pattern of what *must* happen in the system. The two masses in the example above would no longer instance a pattern of *attraction*, only of *moving closer together*. As I discuss in the next few sections, philosophers disagree about how to ontologically characterize whatever it is that realizes the modal aspect of causal patterns (which they usually refer to as 'causal structures' instead; I will follow their usage from now on).

2.3.2 Multi-Perspectival Realism and Causal Structure

A new type of "multi-perspectival realism" (to use Wimsatt's 2007 terminology) has been emerging that can offer us guidance to finding the right metaphysical account of mechanistic causation given the desiderata of Section 2.2. Sandra Mitchell, for example, writes:

> I will argue for a pluralist-realist approach to ontology, which suggests not that there are multiple worlds, but that there are multiple correct ways to parse our world, individuating a variety of objects and processes that reflect both causal structures and our interests. (2009, p. 13; see also Glennan, 2017, p. 93)

Note that Mitchell takes the value of the approach to lie in the fact that it reflects our interests on the one hand, and causal structures on the other. On her view, the causal structures themselves exist out in the world, independently of human interests and perspectives, but we may adopt more than one perspective or 'ways of parsing' to talk about them. In the words of Gordon Globus, "there *is* a 'reality' independent of perspective, and that 'reality' is *structure*" (1976, p. 282). So there are really two levels of realism here: a perspectival realism about objects and processes, and a non-perspectival realism about causal structures themselves. But the success of this approach hinges on the questions: what are "causal structures"? In what sense are they "real"? What is their metaphysical nature? Can they really do the work Mitchell needs them to do?

Wimsatt seems to make a similar kind of move, arguing that while scientists take multiple perspectives on how the world is carved up,

> this multiple rootedness need not lead to 'anything goes' perspectival relativism, or an anti-naturalist worship of common sense, experience, or language. It yields a kind of multi-perspectival realism anchored in the heterogeneity of 'piecewise' complementary approaches common in biology and the study of complex systems. (2007, p. 12)

Like Mitchell, Wimsatt appeals a notion of causal structure that underwrites such perspectives:

> Ontologically, one could take the primary working matter of the world to be causal relationships, which are connected to one another in a variety of ways—and together make up patterns of causal networks. … These networks should be viewed as a sort of bulk causal matter—an undifferentiated tissue of causal structures … Under some conditions, they are so richly connected that neither perspectives nor levels seem to capture their organization, and for this condition, I have coined the term "causal thickets." (1994, p. 220)

The question, again, is: How are we to characterize this "bulk causal matter" or causal-structural "tissue"? "Thicket"… of what?

2.3.3 Causal Structure as Laws

The traditional way that philosophers have metaphysically characterized causal structures is by positing laws as something ontologically free-standing.[15] Historians such as John Henry have traced this tradition back to Descartes. According to Henry (2004), it was Descartes who was primarily responsible for effecting a shift from the conception of causality as resulting from the intrinsic nature of things, as on the Aristotelian view, to a conception of causality as resulting from laws that are external to physical objects. Henry argues that before Descartes, the notion of a 'natural law' was primarily only invoked to refer to regularity in nature; such usages of 'law' were only descriptive and not explanatory. However, "laws" in

[15] Salmon, who played an important role in the development of the New Mechanist philosophy, did not see causal structures as identical with laws, but instead as 'governed' by them (1984, p. 132).

themselves became central to explanation for Descartes because he considered causal efficacy itself to be located within laws, rather than merely described by them.

Funkenstein (1986) argues that Descartes rejected the idea of physical objects having causal efficacy as part of their intrinsic nature because this was incompatible with his epistemology. Specifically, Descartes's "skeptical analysis of sense perceptions" led to the result that

> "matter" (the object of sense perception) is first and foremost extension, for extension is the only determination of matter perceived "clearly and distinctly." Mathematical relations (and geometry, for Descartes, is throughout quantifiable) constitute all that is known and all that can be known about matter. (Funkenstein, 1986, p. 184)

As a result, matter for Descartes was causally inert in itself. Henry argues that for Descartes, the idea of laws themselves being explanatory, which was novel at the time, was likely inspired by his background in mathematics. But in order to explain physical events, they also had to be causal. At the time of Descartes, the only intuitively plausible alternative to physical matter having intrinsic causal powers was to conceive of things as instead being caused by God. Henry argues that the idea of inanimate bodies being independently capable of "obeying" laws of nature would have been seen as an "awkward inherent implication" of Descartes's position, had he not attributed the role of executing such laws to God.

Eventually, however, Henry's argument continues, people's intuitions shifted so that the idea of inert matter operating according to laws became the conventional and default way of looking at physics. Boyle and Newton still conceived of natural laws as ultimately dependent on God for their execution, but by the nineteenth century, the idea of laws as being ontologically fundamental and free-standing, and as being the ultimate source of causal efficacy (or "self-executing," as it were), became commonplace. This was reinforced, of course, by the success of Newtonian mechanics, in which laws were taken to play a central explanatory role. Newton, who shared Descartes's goal of creating a mathematical system of mechanics, followed Descartes in rejecting the idea of causal powers that inhered within physical objects, and

adopted a view of matter as causally inert and governed by laws. The Cartesian conception of a universe of physical objects causally directed by laws that were extrinsic to those objects continued as the dominant metaphysical picture for philosophers well into the twentieth century. On this picture, matter has no intrinsic causal structure of its own; causal explanations must therefore refer to the laws.

It is important to note at this point that although Newtonian mechanics speaks of "laws," it is not committed to any particular metaphysical picture. From here on, I will use 'Cartesian laws' to refer to the metaphysical conception of laws as universal and extrinsic to matter, and 'Newtonian laws' to refer merely to equations of motion as formulated by Newtonian mechanics (abstracted from any given metaphysical interpretation).

2.3.4 Causal Structures in Analytical Mechanics: Constraints

The Newtonian formulation of mechanics is the one that philosophers generally associate with "classical mechanics." But later formulations, now referred to as analytical mechanics—such as that of Joseph-Louis Lagrange, introduced in the eighteenth century, and that of William Rowan Hamilton, introduced in the nineteenth—are different from Newtonian mechanics in important ways that yield a different picture of the nature of causal structure. Before going into these, it is necessary to introduce some further terminology.

> **Constraint.** A constraint is a limitation on how a system can change. Constraints can be *local* or *non-local*. Consider a very simple physical system that consists of a particle inside a box, and another particle outside the box. The box instances two different constraints: it constrains the first particle from moving out of the region occupied by the box, and it constrains the other one from entering that region. The box is a *local* constraint: it only affects particles in the vicinity of the box's region. The system as a whole may also be subject to the constraint that

something in motion will stay in motion and something at rest will stay at rest (Newton's "first law of motion"). Such *non-local* constraints are less likely to be referred to as "constraints" (and more likely to be referred to as "laws") because their scope is not spatially limited or context-dependent in the way the box constraint is.

Motion equation. A motion equation is a mathematical description of a non-local constraint, or a constraint that is treated as non-local by the coordinate system in question. It describes causal relations that will hold throughout the space in question, or the entire domain over which the coordinate system in question is defined.

Constraint equation. A constraint equation is a mathematical description of a local constraint. For example, it might describe a relation that holds between two specific particles that are bonded together, by equating the distance between their position variables with a constant. Other particles in the system may not be subject to the same constraint.

Coordinate system. A coordinate system is a set of dimensions along which variables in the motion and constraint equations may vary.

Cartesian coordinate system. A coordinate system in which the values of variables each represent points along scalar orthogonal spatial dimensions. Motion equations defined in terms of Cartesian coordinates will hold throughout the entire space in question.

Generalized coordinate system. A coordinate system in which the variables may represent something other than the scalar orthogonal spatial dimensions of the space in question. For example, the motion of a pendulum might be expressed by means of a single scalar position variable that represents the rotation angle of the pendulum arm, or the position along the circular arc of its

path. Generalized coordinate systems are often defined only for specific regions of a given space (for example, the pendulum arc position variable is only defined within the range of possible motion for the pendulum).

Generalized coordinate systems may also be used to treat multiple distinct objects as defined in one coordinate system as single, whole objects in the generalized coordinate system. For example, if points in the Cartesian coordinate system represent positions of particles (as they do in Newtonian mechanics), a generalized coordinate system might be adopted instead in which points represent positions of rigid objects consisting of multiple particles.

Holonomic constraints. In some situations, a system described by a combination of motion and constraint equations that are defined in terms of Cartesian coordinates can be redescribed in terms of a set of motion equations defined in terms of a different ("generalized") system of coordinates. This is done by means of an analytical method that requires being able to integrate the constraint equations. Such "integrable constraint equations" are sometimes referred to as "holonomic constraint equations." Correspondingly, we might call the constraints that such equations mathematically describe "holonomic constraints." The importance of holonomic constraints is that they make it possible to describe the constrained system in terms of motion equations alone in the generalized coordinate system. The simplest example is a system of two particles that are bonded together into a larger, rigid whole. The system can be redescribed in a new coordinate system where a point represents the location of the rigid whole, instead of one of the two particles. This enables a reduction in the number of variables needed to describe the system.

Non-holonomic constraints. Non-holonomic constraint equations are constraint equations that cannot be integrated, so that it is not possible to analytically derive

a new description of the system in terms of only motion equations in a generalized coordinate system. The system consisting of a box and two particles was an example of this. Because the box constrains only the position of particles while leaving the velocity free to vary independently, no generalized coordinate system can be derived that could describe the same behavior in motion equations alone.

With these definitions on the table, the two key differences between Newtonian mechanics and analytical mechanics that are important for present purposes can be stated as follows:

- Whereas the Newtonian formulation relies only on motion equations, analytical mechanics introduces constraint equations.
- Whereas the Newtonian formulation always expresses motion using Cartesian coordinates, analytical mechanical methods often involve switching to a new coordinate system (referred to as generalized coordinates) that is tailored to the problem at hand.

Though it remains possible (in principle) to describe any classical system using the Newtonian formulation, the innovations of analytical mechanics allow many types of problems to be expressed in a far more economical and practical manner. This is because there are many types of patterns and structure that can be represented and exploited by analytical mechanics that cannot be represented and exploited by Newtonian mechanics. Take for example a rigid object composed of many particles. In Newtonian mechanics, the system will require six variables for each particle (location and velocity in three dimensions), and a separate equation for each particle. This becomes quite unwieldy for systems with a large number of particles. Instead, analytical methods allow reduction of the entire system to a single equation by adopting a “generalized” coordinate system that might consist of only eight variables,

representing degrees of freedom (three each for location of the center of mass and velocity, and two for angles of rotation about the center of mass) for the rigid object as a whole.[16]

The transformation from Cartesian coordinates to generalized coordinates begins with the addition of constraint equations. Whereas the motion equations describe how the coordinate variables will change over time, constraint equations describe limitations on how they can change. In cases where the constraint equations are not too complex (for example, when they do not have to be expressed using differential equations, in which cases the constraints are referred to as holonomic), analytical methods may be applied to the motion and constraint equations to derive a new description of the system in terms of motion equations in a generalized coordinate system. It is important to realize, however, that in such cases, the constraints have not ceased to exist; they are merely implicit and are now incorporated into the new "laws of motion" of the system. The new "laws of motion," that is, the equations that describe the system, are not laws in Descartes's sense; they are equations that only apply locally to the system that has been so redescribed. In many other cases, however, constraint equations cannot be eliminated by analytical means. I will refer to the former cases as 'holonomic systems', and the latter as 'non-holonomic systems'.[17] Holonomic systems will include cases like rigid objects; non-holonomic cases will include systems with complex, machine-like dynamics, notably including protein molecules, which make up the building blocks and machinery of all biological systems (Pattee, 1973a).

The need for formulations other than that of Newton arises because there is a great deal of causal structure that Newtonian equations do not describe; not all structure is

[16] For recent textbooks that deal with these topics, see Goldstein, Poole, and Safko (2002) and Taylor (2005).

[17] Technically this is not quite accurate because analytical methods do exist for producing constraint-free formulations of some relatively simple non-holonomic systems (specifically, those that fall under the scope of D'Alembert's principle), but this technicality will be of no consequence in what follows, since biological systems are of much greater dynamical complexity (Pattee, 1973a).

microstructure (Mitchell, 2012).[18] Much of this causal structure is confined locally; this is why it is often useful to adopt problem-specific constraint equations or coordinate systems. The Newtonian formulation only exploits facts about causal structure that hold true in all parts of space and time, for all kinds of (classical) dynamical systems. This is why it expresses all problems in the same, Cartesian, coordinate system. Dynamical systems include a great deal of structure that is not described by Newtonian mechanics, and non-holonomic systems have dynamics that cannot be described without reference to constraints. What are the metaphysical implications?

2.3.5 A Metaphysics Inspired by Analytical Mechanics: Constraints as Ontologically Primitive Modal Structures

In order to account metaphysically for the fact that Newtonian motion equations capture causal structure on the particle level, and generalized motion equations and constraint equations capture larger and higher-order causal patterns, there are three options:

1. Treat both motion equations and constraint equations as true in virtue of Cartesian laws, placing causal efficacy within ontologically robust, standalone entities that are extrinsic to all objects and dynamical systems.
2. Treat motion equations as true in virtue of ontologically robust, extrinsic entities (Cartesian laws), and constraint equations as true in virtue of the intrinsic causal structure of constrained systems (so that causal efficacy is partly extrinsic to objects in the world and partly intrinsic).

[18] Further, not all causal structure is reducible to microstructure (Wimsatt, 1994; Mitchell, 2012).

3. Treat both motion equations and constraint equations as true in virtue of causal structure that is intrinsic to the particles, objects, and dynamical systems themselves.

In my view, Occam's razor and other considerations clearly favor the third option. The first option would require non-universal extrinsic principles of motion that undergo continual change (as local constraints continually change, and dynamical systems become describable in terms of different kinds of constraint equations). Most of the appeal of regarding principles of motion as extrinsic to objects is supposed to derive from their being the universal and unchanging anchors of physical explanation. Option one would defeat this basic appeal. Option one also leaves the dynamics of non-holonomic systems unaccounted for. Option two locates causal efficacy partly within extrinsic laws and partly within intrinsic natures. Some advocates of 'strong emergence' have proposed such a metaphysics.[19] But given that some systems (for example, holonomic systems) could equally be described either in terms of motion equations alone or in terms of a combination of motion and constraint equations, it does not make sense to argue that there is a single, objective partitioning in nature between extrinsic and intrinsic causal principles of motion.

The preferable option is the third, that is, the metaphysical idea that causal structure in general is something local to and inherent within mechanical systems, waiting to be discovered and exploited by means of various choices of motion equations, constraint equations, and coordinate transformations. Systems can be described by these means because they intrinsically possess systematic limitations on how they can change. On option 3, causal structure refers precisely to such intrinsic systematic limitations and is ontologically primitive (that is, it is not grounded in an extrinsic principle of change like Cartesian laws). Like Ross, Ladyman, and Spurrett (2007), and in order to mark the departure from the Cartesian

[19] For critical discussion of such views, see Klee (1984).

metaphysics of laws, I will refer to such ontologically primitive, intrinsic limitations as ‘constraints’. On this view, constraints are more than mere regularities; in the words of Mumford (2004), constraints are “modally loaded.” They may be thought of as modal patterns. Often, patterns are conceived in philosophy as nothing more than non-modal regularities. But constraints are more than just occurrent regularities; constraints in a dynamical system pertain to what might happen. They are the modal facts about a dynamical system, the truthmakers for dynamical equations and modal causal claims.

This conception of causal structure is cognate with French’s “modal structuralism” which “takes the structure to be ‘inherently’ modal” (2014, p. 263) and Esfeld’s (2009) conception of “causal structures.” French (2006, p. 183) argues that there is nothing inherent to the concept of a structure that makes it any less causal than other ontological categories like properties. French also draws inspiration from Maudlin (2007) to make a positive argument for his modal structuralism. Maudlin argues that

> nothing in *scientific practice* suggests that one ought to try to reduce fundamental laws to anything else. … The practice of science, I suggest, takes fundamental laws of nature as further unanalyzable primitives. As philosophers, I think we can do no better than to follow this lead. (2007, p. 105)

French carries this reasoning further, arguing that the practice of physics shows that physicists treat laws as structures that are causal, unanalyzable, and primitively modal (2014, p. 298). As I have argued, however, looking at scientific practice leads us not to the Cartesian conception of causal structure as extrinsic laws, but instead to a conception of causal structure as intrinsic constraint. This argument is only strengthened when we realize that quantum mechanics is itself an offshoot of analytical mechanics, not Newtonian mechanics (Dirac, 1958, p. 114).

The present view is a rejection of the Cartesian metaphysical account of laws. But it is not necessarily a rejection of the idea of a “law” *tout court*. Ashby (1956, p. 130) characterized laws as merely a special case of physical constraints that apply universally. Similarly, Glennan (2017, pp. 3 & 44–46) characterizes laws as generalizations about local, intrinsic causal

structures. But these need not be universal. Cartwright (1999), for example, argues that empirical evidence does not support universal generalizations, and that what are called “laws” are generally only true within the scope of local “nomological machines.”

These considerations support the view that mechanical systems inherently contain a “thicket” of constraints. It is our choice how to represent these constraints, that is, how to “parse our world” (to use Mitchell’s phrase) and individuate individuals and processes—the task accomplished in analytical mechanics by choosing which coordinate system and/or constraint equations to adopt. A constraints-based metaphysics provides a more comprehensive way to understand causal structure that is rooted in scientific practice.

2.4 Constraints and Mechanistic Causal Powers

I now return to two questions raised but not yet answered from Section 2.2: What could it mean for something to be part of the actual, intrinsic nature of an object, but in a non-manifested state? And can this be understood in a way that is grounded in the details of scientific practice?

2.4.1 Inter- versus Intra-Perspectival Categories

As physicist and theoretical biologist Howard Pattee explained, constraint is really an inter-perspective concept, rather than an intra-perspective one. This is because the concept of a constraint is really the concept of a set of possibilities that is reduced to a smaller number of possibilities. Describing something as a constraint implies at least two ways of carving the world. This can be seen in the example of the rigid object composed by particles, which might allow a change in coordinate systems. By referring to the rigid object as constrained, one

makes implicit reference to the fact that without the constraint, the system would consist of particles with many more degrees of freedom:

> constraints must be defined by different descriptive levels. … Why are these necessarily two-level processes? Why are two distinct descriptions necessary? Because we cannot speak of an event as being both possible and impossible using the same level of description. On the lower, unconstrained level the alternatives must be possible; for if they were impossible then deciding for or against them would be a vacuous process. But on the upper, constrained or controlled level, … some of these alternatives are actually selected, or more precisely, made more probable … (Pattee, 1972, p. 84)

For this reason, writes Pattee, "whenever a physicist adds an equation of constraint to the equations of motion, he is really writing in two languages at the same time" (1973a, p. 98). Constraints require some kind of stuff in order to be instantiated, but their instantiation does not depend on how the underlying stuff is ontologically "carved" into objects, or what Lloyd Morgan called the "substantial gotogetherness" (1927, p. 193) of the stuff. This is a key difference between constraints and powers. Powers emerge once one adopts a world-carving perspective; constraints are the preconditions in the world that underwrite the adoption of such perspectives.

Return to the example of the chaperone and substrate molecules. From one perspective, each is a collection of atoms held together in a certain way by bonds. From another perspective, each is an object constrained to possess a certain shape. Neither of these perspectives is the "right" one; they merely represent distinct "ways of parsing" the world. However, in each case the same underlying, invariant causal structures (whether we conceptualize them as configurations of bonds between atoms or as shapes of objects)—constraints—are determining how the system will behave.

It might be argued that when we abstract away from object-oriented ways of carving up the world, and merely refer to the universe as consisting of an uncarved expanse of "stuff" that is constrained in various ways in various locations, we are still adopting a "perspective." Yes, but this will by definition not be a world-carving perspective. When I say that constraint is an

inter-perspectival ontological category, I mean that it is an ontological category that is independent of any perspective on where the boundaries between objects (events, processes, and so on) are. It is an ontological category that picks out the full range of real causal patterns in the world which are the candidates for populating such world-carving schemes; any world-carving perspective will selectively isolate a subset of such causal patterns to form the basis of its parcellation into objects, events, processes, and so on.

The key difference between powers and constraints, namely that powers are applicable to object-oriented ways of carving up the world, whereas constraints are inter-perspectival, demonstrates why it is necessary to have both categories. Here it is helpful to recall MDC's claim that causal powers are "substantivalist." In the framework of mechanistic explanation, any power instance is the power of some object (that is, of a mechanism or part of a mechanism); if the object goes out of existence (that is, its identity conditions no longer obtain), its powers do as well. Unlike causal powers, constraints don't go away when you change ontological perspectives.[20]

2.4.2 Mechanistic Causal Powers are Grounded by Constraints

My main suggestion in this chapter is that we adopt a multi-perspectival realist approach to understanding mechanistic causation. Again, multi-perspectivalism involves two levels of realism: a perspectival realism about objects and processes, and a non-perspectival realism about causal structures themselves.[21] Correspondingly, a metaphysical account of

[20] It should be noted that while most ways of conceiving causal powers tie them to objects or processes, this is not true of all ways of conceiving causal powers. For example, Marmodoro (2017) makes a distinction between structural powers and substantial powers; her 'substantial power' corresponds to my use of 'power', and her 'structural power' is closer to what I am calling 'constraint'.

[21] Ross, Ladyman, and Collier provide reasons why "our local actual (physical) circumstances [might] be such that constructing individuals is necessary for tracking some extra-representational real patterns" (2007, p. 245). What they refer to as "locally dynamic real patterns" (e.g., 2007, p. 252) seems to be the

mechanistic causation must involve perspectival realism about causal powers, and non-perspectival realism about constraints. Mechanisms, being physical systems, are constituted by physical stuff; as such, constitution is a relation that can transcend world-carving perspectives. Then what is the relation between constraints and causal powers? The relation cannot be identity, since something that is real relative to a perspective will not have the same identity conditions as something that is real independently of perspective. Instead, the relation is a certain kind of ontological dependence: the constraints, that is, the perspective-independent causal structures, are what make it possible to usefully adopt a causal powers perspective. When we adopt an object-oriented perspective, the way that the stuff constituting an object is dynamically constrained allows us to talk instead of powers that the object has.[22] Constraints are the truthmakers for modal facts about the dynamics of a system; when looked at from a perspective that parcels the system into objects and properties, these modal facts map onto dispositional facts about such objects and properties. In other words, they serve as the intrinsic and actual grounding of perspectival facts about what kinds of behaviors will manifest under certain conditions.

The causal powers approach seemed like a promising way to account for mechanistic causation until we saw that it could not meet two of the desiderata in a way that practice-oriented philosophers of science would consider fully satisfactory: scientific validity/non-mysteriousness, and the perspectival nature of mechanisms. However, if we can account for causal powers ontologically in terms of constraints, we will have an account of causal powers that is well-grounded in the details of scientific practice; Paul C. W. Davies writes that "in almost every branch of physics, with the possible exception of elementary particle physics,

---

same as what I am calling "constraints." Later in the same book, Ross, Ladyman, and Spurrett (2007, p. 288) use the word 'constraint'.

[22] Juarrero (1999, pp. 131–132) and Moreno and Mossio (2015, p. 51) have also proposed that constraints can ground causal powers. Cartwright and Pemberton (2013, p. 96) reverse the order of priority and hold that "constrainings" result from causal powers.

constraints play a crucial role" (1989, p. 104). Further, whereas resorting to causal powers alone does not provide a non-perspectival grounding of mechanistic causation, constraints offer a way to account for the underlying causal structure that underwrites causal powers perspectives.

The present account of mechanistic causation will not be complete, however, until it is explained how constraints can ground the other aspects of causal powers that allow them to satisfy the remaining three desiderata: intrinsicness, productivity, and directionality.

2.4.3 Intrinsicness and Constraints

Unlike laws, constraints are local to the system that possesses them. The local, intrinsic aspect of constraints has been emphasized by a number of authors who have applied the concept of a physical constraint to a wide range of contexts.[23] It is important to note here that sometimes authors have distinguished between "intrinsic" and "extrinsic" constraints, or "internal" and "external" constraints. For example, "constraints may be external owing to the environment interacting with the system. Or such constraints may arise internally within the system owing to the collective effects of its constituents or the evolving dynamics" (Bishop, 2012, p. 5). Deacon uses the words 'intrinsic' and 'extrinsic': "constraints can originate intrinsic or extrinsic to the system that is thereby constrained" (2011, p. 549). The distinction between intrinsic and extrinsic constraints is important in thermodynamics, for example (Nicolis & Prigogine, 1977).

However, these are just different ways of talking about *metaphysically* intrinsic constraints. Hooker (2013, p. 760) argues that any system that is described in terms of external

[23] For example, situation theory (Barwise & Perry, 1983, p. 98), theoretical biology (Pattee, 1982, p. 176), cybernetics (Ashby, 1956, p. 131), information theory (Deacon, 2007, pp. 127ff.), behavioral science (Kelso, 1995), ecological psychology (Greeno, 1994), and Gestalt theory (Köhler, 1922/1967, pp. 61–62 & 68–69).

constraints between subsystems can be re-described in terms of the internal constraints of the larger containing system. But he argues that this does not in any way diminish the reality of the constraints themselves; it is only the characterization of them as internal versus external that is perspectival. For example, in the case of Rayleigh–Bénard convection, the Bénard cells that form are themselves internal constraints of the boiling water that result from the external constraints of the water's metal container and the heat source. Nothing stops us from describing it as a single system that contains the heat source, metal container, and the Bénard cells themselves as internal constraints. But this does not show that constraints are not local or intrinsic to dynamical systems. It shows, instead, that constraints remain regardless of the perspective we take on the portion of physical stuff that is constrained. We can talk about that stuff as being a single object, or a set of objects, or a portion of a larger object; in all of these cases, we are talking about the same locally constrained stuff in different ways. One is not more fundamental than the other.

2.4.4 Constraints and Productiveness

In Section 2.3.5, I defined constraints as ontologically primitive intrinsic limitations on how dynamical systems can change. But how is the idea of something limiting the possibilities for change of a system compatible with its producing the resulting changes? In one sense, of course, constraints are limiting, and reduce the degrees of freedom of a system. But in another sense, constraints are enabling, and they shape and define the types of behaviors a system will have, as Cliff Hooker explains:

> … constraints can at the same time also be enabling, they can provide access to new states unavailable to the unconstrained system: equivalently, by coordinatedly decreasing degrees of freedom they provide access to dynamical trajectories inaccessible to the unconstrained system. … a skeleton is a disabling constraint, for example limiting the movements of limbs (cf. an octopus), but by providing a jointed frame of rigid components for muscular attachments it also

> acts to enable a huge range of articulated motions and leverages (2013, p. 761; see also Pattee, 1973a)

To see how productivity results from constraint, consider what a dynamical system would be like if it were absolutely unconstrained. From one moment to the next, all configurations or state-transitions would be equally probable. It would not be possible to make any kind of predictions about the system (or, for that matter, to take an object-oriented perspective on events going on in the system). All predictability requires redundancy, which is nothing more or less than a way that a system is constrained. Whenever we can isolate systematic or predictable behaviors or forces within a system, that reflects a way in which the system is organized. Therefore, writes Ashby, "the presence of 'organization' between variables is equivalent to the existence of a constraint in the product-space of the possibilities" (1962, p. 257). When we put this fact together with the recognition that productive relations within mechanisms exist in virtue of the causal organization that mechanisms possess (Glennan, 2017, p. 23)—causal organization which underwrites and makes possible the taking of any mechanistic perspective—the sense in which productive relations within biological systems derive from how such systems are constrained becomes clear.

2.4.5 Constraints and Directionality

Recall Bechtel's and Glennan's point above that in most mechanistic operations, there is both something that acts and something that is acted upon. A mechanistic perspective will therefore usually be one that posits directional causal powers: the operation will involve an active component and a passive component, and the effects will be explained in terms of active powers of the active component and passive powers of the passive component. An account of how constraints ground causal powers should provide a story about how directional powers are grounded.

In Section 2.4.3, I considered the fact that a dynamical system may contain subsystems, and that we can then distinguish between the subsystem's internal and external constraints. Its internal constraints are instantiated by the stuff that constitutes the subsystem; the external constraints are instantiated by stuff that is external to it. Since a passive power is the ability of something to be causally affected in a certain way by something external to it, a passive power can only be instantiated by an object constituted by stuff that is both internally and externally constrained in certain ways. Specifically, I consider passive powers to emerge from a kind of second-order constraint:

> A **passive causal power** exists when a system within a larger system is internally constrained in such a way as *to be externally constrained* under certain conditions.

Similarly, active causal powers (of the kind that result in causal effects within external objects) also emerge from a kind of second-order constraint:

> An **active causal power** exists when a system within a larger system is internally constrained in such a way as *to externally constrain* under certain conditions.

We can now revisit the example of the chaperone molecule performing the operation of folding a substrate protein. The chaperone plays an active role in the process, since it changes the structure of the folding substrate without itself being changed in the relevant way by that process. The chaperone has the active causal power to fold the substrate: it has an internal bond structure that will allow it to externally constrain the structure of the substrate after the substrate bonds to the chaperone. After folding is completed, the bonding structure of the substrate has changed, but the chaperone returns to its original configuration. The substrate had the passive causal power to be folded: it had an internal bond structure that would allow it to be recognized by the chaperone and to be folded by it.

Keep in mind that there is nothing ontologically profligate about countenancing second-order constraints; constraints are merely a type of structure, and any account of structure will

include higher-order structures. There are of course many orders of structure within biological systems, and this is key to Wimsatt's notion of "causal thickets." Because these higher orders of structure are indispensable to biological explanation, as well as to the other "sciences of complexity" (Stein, 1989), we should acknowledge that primitive modality may also accompany these higher orders.

2.5 Conclusion

In this chapter, I have argued that extant accounts of the metaphysics of mechanistic causation do not succeed because they fail to satisfy five key desiderata that have emerged in recent New Mechanist literature. I offered a novel multi-perspectivalist approach to understanding mechanistic causation which incorporates two levels of realism: a perspectival realism about objects (mechanisms and their parts) and their causal powers, and a non-perspectival realism about the causal structures that underwrite these perspectives. I further showed how the conception of non-perspectivally real causal structures can be cashed out in terms of a concept that has long been familiar to science but that has been mostly neglected by philosophers: constraints. Finally, I demonstrated how a constraints-based account has the resources to satisfy the five desiderata, and to provide a non-perspectival grounding for mechanistic causation.

Acknowledgement

Chapter 2, in full, is a slightly expanded version of the material as it appears in "Mechanistic Causation and Constraints: Perspectival Parts and Powers, Non-Perspectival Modal Patterns," *British Journal for the Philosophy of Science*, forthcoming. The dissertation author was the sole author of this paper.

Chapter 3
Internal Perspectivalism:
The Solution to Generality Problems about Proper Function and Natural Norms

3.1 Introduction

In a wide variety of contexts, scientists and philosophers take for granted that it is meaningful to refer to or inquire into *the* biological function of a given trait, property, or part (from now on I will just say 'trait' for sake of brevity) of an organism or biological system. The idea here is that although a given trait may be capable of performing any number of functions—in the sense that there are various causal interactions it may enter into, and hence various causal roles it can play in those interactions—there might only be one function that it *has* (its *proper function*).[24] The distinction between properly functioning versus malfunctioning traits is extremely important to philosophy because it has been thought to provide resources for resolving a number of philosophical debates. For example, proper function has figured centrally in philosophical accounts of representation (Dretske, 1988), computation (Piccinini, 2015, p. 11), action and agency (Burge, 2009), mechanisms and mechanistic explanation in biology (Garson, 2013), epistemic norms (Sullivan-Bissett, 2017), and even value as such (Foot,

[24] This is a broader use of 'proper function' than is sometimes used; often, 'proper function' is used to refer specifically to the etiological conception of function defended by philosophers such as Karen Neander (1991). On the present use of 'proper function', the etiological conception is only one of many conceptions of proper function. Often, 'teleological function' is used to mean what I intend by 'proper function' (e.g., Maley & Piccinini, 2017). Not all accounts of proper function (as I am using the term) provide for a notion of malfunction, as I detail below. It should also be noted that, at least on some views, traits can have multiple proper functions; for example, it might be said that the hind legs of turtles have the proper functions both of locomotion and of excavation (Preston, 1998).

2001), among many other projects falling under the heading of "naturalizing normativity." Proper function is an equally important notion for biologists. Their investigations are often inquiries into what *the* function of a given trait is, or into what *the* function of parts that contribute to produce a larger phenomenon or behavior are, or what *the* function of certain activities or developmental processes (I will also include these under the umbrella of 'trait') are in the life cycle of given organisms.

In order to provide an account of proper functions, philosophers need to answer two questions:

> **Demarcation Question:** What is the difference between a properly functioning trait and a malfunctioning one?
>
> **Grounding Question:** What *makes it the case that* traits have functions rather than merely accidental causal dispositions?[25]

The first question is about the traits themselves. The second question is about the situatedness of traits that makes them have any normative or functional status at all. I will divide the approaches philosophers have taken to answering these questions into two broad categories: perspectival and non-perspectival. The difference is that on perspectival approaches, proper function ascription is irreducibly relative to a perspective.

In this chapter, I raise a problem for non-perspectival accounts in philosophy of biology that is analogous to the *generality problem* that has been raised against process reliabilism in epistemology (Section 3.2). I argue that this should not be surprising: Both types of account explain a normative status (epistemic justification of beliefs in one case, proper function of traits

[25] It is important to keep in mind that these are *metaphysical* questions; i.e., they are questions about the nature of proper functions themselves, rather than epistemological questions about how humans can/should go about studying them. In this chapter, I do not address questions about how humans discover, test/confirm hypotheses about, or reason about proper functions. Though I don't offer an account of the role that proper function plays in scientific explanation, I do consider the *fact that* proper functions play an explanatory role for biologists as a desideratum that can rule out certain metaphysical accounts of proper function (see Section 3.4).

in the other) by reference to the type a certain token (third column of Table 3.1) falls under, but such tokens fall under many such types of varying generality, and neither account offers a way to tell us which is the relevant type in given cases (Section 3.3). Perspectival accounts of proper function[26] do not have this problem, since the perspective that a functional attribution is relativized to will involve a univocal interpretation of the situation that disambiguates between the problematic multitude of types.

I further divide perspectival accounts into two categories: external perspectivalism and internal perspectivalism. Up to now, discussions of perspectivalism have tended only to focus on external perspectivalism (e.g., Massimi, 2018b). On an external perspectivalist account, proper function within a given system depends on the perspective taken by an observer that is external to the system in question (such as a scientist who is investigating the system). Such accounts have not been as popular among philosophers because they are sometimes not considered to be fully naturalistic and have other well-known problems (Section 3.4). On an internal perspectivalist account, by contrast, the proper function of a trait within a given system depends on the perspective had by the *system itself*. Unlike non-perspectival accounts, internal perspectivalism does not succumb to generality problems. But unlike external perspectivalism, internal perspectivalism can provide a fully naturalistic, mind-independent grounding of proper function and natural norms. Accordingly, I will defend a novel internal perspectivalist account in this chapter, according to which what counts as the proper function of a trait is a matter of the *de facto* perspective that the biological system, itself, possesses on what counts as proper functioning for that trait (Sections 3.5–3.7).

My attribution of perspectives to biological systems is intended to be neither metaphorical nor anthropomorphic: I do not mean to imply that such systems thereby must

[26] As far as I know, no perspectival version of process reliabilism has been offered, perhaps because process reliabilists tend to be epistemic externalists.

possess agency, cognition, intentions, concepts, or mental or psychological states.[27] The notion of a perspective to be developed here (in Sections 3.5 and 3.6) is much thinner, depending only on minimal recognition and response capabilities by means of which even "simple" organisms without nervous systems, and subsystems within such organisms, are selectively sensitive to details of their outer and inner environment at a fixed level of generality. Such systems provide the grounding for norms of performance when they internally enforce their own standard of (i.e., their own perspective on) what constitutes proper functioning or malfunctioning. Since they operate with a fixed, determinate level of generality, such systems provide the basis for an account of proper function that is immune to generality problems.

Importantly, the sort of generality problem I will raise for theories of proper function is not merely one of fuzzy boundaries that result in mildly puzzling cases or sorities paradoxes.[28] An account that suffers from generality problems not only fails to give an exact answer; it fails even to give you the means to know when you are anywhere near the right answer, as I demonstrate in the next section.

3.2 The Generality Problem for Process Reliabilism

Generality problems have been most thoroughly discussed in debates about the nature of epistemic justification. Process reliabilism is a theory of epistemic justification stating that

> The justificational status of a belief is a function of the reliability of the process or processes that cause it, where (as a first approximation) reliability consists in the tendency of a process to produce beliefs that are true rather than false. (Goldman, 1979, p. 10)

[27] My internal perspectivalism is therefore very different from that of Sinnott, who argued that "biological organization … and psychical activity … *are fundamentally the same thing*" (1961, p. 48). Pattee, whose work partly inspired the present view, might also be read as an internal perspectivalist about biological function (e.g., 1970, p. 130; 1982).

[28] See Neander (1995, p. 113) for a useful breakdown of different kinds of function indeterminacy.

Process reliabilism attempts to answer epistemic analogs of what I have referred to as the Demarcation and Grounding questions: it attempts to say what makes a justified belief different from an unjustified belief (a justified belief must have been formed by a *reliable* type of process), and it attempts to tell us what makes it the case that beliefs have any justificational status in the first place (the purported fact that any given belief will have been formed by a process that has some determinate level of reliability).

Conee and Feldman (1998) argue that belief-forming process tokens will always fall under a number of different types of varying generality and of varying reliability. For example, suppose I look out my window and form the belief that my friend Sharon is walking on the sidewalk outside. Suppose further that the process by which I formed my belief counts as a i) visual process at night, ii) visual process at night in a well-lit area (there is a street light nearby), and iii) visual process on a foggy night. We might characterize my belief as having been formed by any of these process types, among others, depending on the level of generality at which we choose to individuate such processes. But note that some of these belief-forming process types (BFPTs) are reliable, and some are not. Process reliabilism will therefore give wildly differing answers depending on the level of generality at which we individuate processes. Without resources for specifying the relevant level of BFPT generality, such an account fails even to give you the means to know when you are anywhere near the right answer. It therefore does not provide a satisfactory answer to the grounding question: it does not tell us which non-normative facts make it the case that beliefs are justified or unjustified.

As Michael Bishop put it, “Without a principled solution to the generality problem, the reliabilist can always start with the intuitively correct judgment about a belief’s justificatory status and then cherry pick a BFPT that yields that judgment,” (2010, p. 287). Conee and Feldman argue that

> Given the multiplicity of belief-forming process types and their variations in reliability, it is easy to make *ad hoc* case-by-case selections of types that match

> our intuitions. But case-by-case selections of relevant types does not constitute working out a reliabilist theory of justification. (1998, pp. 3–4)

In the intervening years, solutions have been proposed, but for my purposes what matters is that it continues to be recognized as a serious problem. In the next section I argue that a parallel and equally serious problem afflicts each of the three major non-perspectivalist theories of proper function.

3.3 How Generality Problems Arise for Non-Perspectival Theories of Function

Space does not here permit an exhaustive critique of non-perspectivalist theories, but I will briefly indicate a few ways that generality problems can arise in three prominent non-perspectivalist approaches: etiological, systemic causal role, and self-maintenance-based accounts. Generality problems arise for these theories because they each analyze the proper function of traits in terms of how those traits relate to some related token entity (third column

Table 3.1: Analyses that result in generality problems due to the fact that the token referred to in the analysans (third column) can be considered as falling under multiple types of varying generality.

Name of account	Normative analysandum	Token(s) invoked in analysans that can be typed at multiple levels of generality
Process Reliabilism	Epistemic justification	Process by which belief was formed
Etiological theories	Proper function	Task that the trait was selected for; selective regime
Systemic causal role theories	Proper function	System that the trait is part of; systemic capacity; contribution to system capacity
Self-maintenance-based theories	Proper function	Self-maintenance regime for that trait; organizational class

in Table 3.1) that falls under multiple types. Just as you can reach conflicting judgments about justification by characterizing belief-forming processes as instances of types of varying generality, you can also reach conflicting judgments about proper function by characterizing selection processes, systems, and self-maintenance regimes as instances of types of varying generality, as I will presently demonstrate.

3.3.1 Etiological Approach

According to the most prominent etiological approach, "a token trait's function depends on what traits of the relevant type were selected for" (Neander & Rosenberg, 2012, p. 613).[29] Whether or not trait X is functioning properly is a matter of whether or not trait X can perform the task that it was selected for performing. It is therefore the fact that the trait was selected for a certain task that provides the basis for attributing *proper* functions rather than mere causal ones (an answer to the Grounding question), and a trait functions properly if and only if it currently can perform that task and provide the fitness contribution in question (hence answering the Demarcation question).

Whether or not a trait can perform the task it was selected for performing depends on what task it was selected for performing. The assumption is that on many occasions in the past, the trait helped the organism's ancestors survive and reproduce to a greater extent than conspecifics without the trait in question. In order for these occasions to add up to a definite

[29] I count the etiological approach as non-perspectival because, as many authors have pointed out (e.g., Polányi, 1958, p. 385; Nagel, 1977, pp. 286–287; Nyberg, 2009, p. 187; Okasha, 2009, p. 720), "natural selection" does not involve anything like an intelligent agent in the background applying any particular normative standard. Instead, in any given case, organisms succeed or fail at reproducing, not because of some universal trait that is consistently selected across eons and ecosystems, but because of whatever particular traits happen to be helpful (or "adaptive") in specific environments on specific occasions (Beatty, 1984, pp. 192–193).

function attribution, it must have been beneficial in the same way across those occasions, i.e., by performing the same task.

A generality problem arises because the task at issue can be described at different levels of generality. For example, suppose the auditory system of a preyed-upon mouse population has become especially attuned to a certain frequency band B over a large number of generations, because its main predator (a certain hawk population) makes a distinctive sound in which frequency band B is dominant.[30] Suppose further that over those many generations, due to the nature of their habitat, the mice gained almost no survival or reproductive benefit from detecting noises otherwise. Now suppose a mutation occurs in the hawk population, making it almost silent in frequency band B (producing what we may call "stealthy hawks"; call the non-mutated hawks "noisy hawks"). We may ask, is the auditory system of one of the mice malfunctioning when it fails to detect the stealthy hawk approaching? This depends on how we characterize the task that the auditory system was selected for, i.e., the task that conferred a benefit to its possessors. We may describe this as:

- Detecting predators that are noisy with respect to frequency band B
- Detecting noisy hawks
- Detecting hawks
- Detecting predators
- Detecting sounds in general

If "detecting sounds in general" is the task that the auditory system was selected for, then arguably, it was not malfunctioning. The mouse was generally able to detect sounds, just not those made by the stealthy hawks. Similarly, if "detecting predators that are noisy with respect to frequency band B" is the selected-for task, then it is not malfunctioning. But if the task is

[30] This example is loosely inspired by one given by Walsh (1996), but it is being used to make a different kind of argument.

"detecting hawks" or "detecting predators," then it is indeed malfunctioning. Our intuitions may pull us toward one answer or another, but as was the case with process reliabilism, it will be our intuitions doing the heavy lifting, not the account itself.[31]

The larger fact at issue is that, as Walsh argues, natural selection is "relative to a selective regime" (1996, p. 553), defined as "the total set of abiological and biological (including social, developmental and physiological) factors in the environment of the trait which potentially affect the fitness of individuals with that trait" (1996, p. 564; cf. Bechtel, 1986). Etiological theories of function therefore "face indeterminacy problems, for there are more and less specific descriptions of selective regimes and functional outcomes" (Sterelny, 1995, p. 255). Goode and Griffiths write that "the apparent indeterminacy of etiological functions is a genuine indeterminacy, but a harmless one. Selection processes can be described at more or less abstract theoretical levels, all of which generate genuine, complementary descriptions of etiological function," (1995, p. 107). But as the example above shows, descriptions at differing levels of generality will often generate conflicting, not complementary, function and malfunction ascriptions.

The point here is that the bare historical facts of "selection," by themselves, are insufficient to pick out proper functions and malfunctions in given cases. The facts, by themselves, do not specify the relevant level of generality for making function assignments. To paraphrase Bishop, it is too easy for the etiological theorist to "start with the intuitively correct judgment about a [trait's proper function] and then cherry pick an account of [the organism's selection history] that yields that judgment."[32]

---

[31] Neander addresses this type of problem by arguing that when there are multiple ways to characterize functions, "we should give priority to that description of a trait's function that is the lowest level in the analysis (most mechanistic)," where mechanistic levels are connected by "asymmetrical by-relations" (1995, p. 137). But in the above case, unlike the cases Neander considers, the varying characterizations do not correspond to different mechanistic levels connected by asymmetrical "by-relations."

[32] Enç (2002) enumerates a series of arguments against etiological theories of proper function along similar lines as well.

3.3.2 Systemic Causal Role Approach

Systemic causal role theories of function attribute functions to traits on the basis of the causal contribution that such traits make to the activities of the larger system that they are part of. Some authors (e.g., Craver 2001, 2013) explicitly relativize such accounts to an external perspective; here, I will consider non-perspectival versions of such accounts (external perspectival accounts will be considered in Section 3.4). As well, many authors who advocate for systemic causal role theories explicitly deny that their theory is intended to support a distinction between function and malfunction. On this approach, it is the fact that the trait in question has a causal role to play in the organization of the larger system that answers the Grounding question. For versions of the theory that allow for malfunction, the trait is functioning properly if and only if it is can actually play that causal role (hence answering the Demarcation question).

Davies's (2001) account is the most fully worked out version of the systemic causal role theory, which was earlier put forward in different forms by Nagel (1961) and Cummins (1975). Davies's account is not intended to support a distinction between function and malfunction, but unlike Craver's account, Davies's account is intended to be non-perspectival, i.e., it is not grounded in "our explanatory interests or, better, our explanatory whims" (2001, p. 77). Any systemic causal role account not grounded in an external perspective, however, must confront the most common objection leveled against such accounts, which Davies calls the "promiscuity objection." If there are insufficient restrictions on what counts as a "system", and what counts as a "contribution to" the activities of a system, then almost anything can have almost any function (Millikan, 1989, p. 294; Illari & Williamson, 2011, p. 826).

Davies attempts to meet these challenges by offering the following constraint:

> The sorts of phenomena to which the theory of systemic functions properly applies are those that are *hierarchically organized*. A system is hierarchical if it

> exercises a capacity at one level by virtue of the organized capacities operating at some lower level of organization. (2001, p. 82)

Davies argues that “restricting the theory to hierarchically organized systems provides resources with which to distinguish the functional from the merely causal” (2001, p. 86). But even with this restriction, contributions, systems, and system capacities can be described at multiple levels of generality, leading to multiple conflicting conclusions about whether a function attribution should be made and what it should be.

Consider the function attribution being made in the following passage:

> Human fingers are capable of gripping a cigarette. The fingers, combined with the mouth and respiratory tract, constitute a larger system that has the capacity of smoking a cigarette (call this the “cigarette-smoking system”). Since the fingers' ability to grip the cigarette contributes to the larger system's capacity to smoke the cigarette, gripping cigarettes is a biological function (not a mere accidental capacity) of fingers.

The systemic causal role theorist could simply agree with the conclusion, but the theory now seems dangerously close to counting almost any capacity as a biological function. Here is the beginning of a list of other ways the systemic causal role theorist might respond to the passage:

- Cigarette-smoking is a capacity realized at the same level as the fingers, since there is no such higher-level system as the “cigarette-smoking system”. Since the hierarchy constraint is not being fulfilled, the fingers should not be said to be exercising a biological function in this case (no function is being exercised).
- The larger system capacity has been identified incorrectly. It is more accurate to say that the fingers are contributing to the musculoskeletal system’s capacity to position an object (the function is therefore “to hold an object being positioned”).
- The larger system capacity has been identified incorrectly. It is more accurate to say that the fingers are contributing to the whole-organism capacity to

consume an object (the function is therefore “to grip an object being consumed”).

- The task performed by the fingers has been identified incorrectly. The contribution that the fingers are making is to apply forces to a cigarette at different points and in opposing directions. The larger systemic capacity is the hand’s ability to hold a cigarette steadily in place (the function is therefore “to apply forces to a cigarette at different points and in opposing directions”).

… and so on.

These statements are not equivalent; some of the functions could be lost without losing others.[33] The systemic causal theorist will probably have the intuition that some of these statements should be preferred over the others, and perhaps even that one of them is the sole correct one. But we are now at the same point we were at with process reliabilism in Section 3.2. All you have to do is find the right level of generality at which to describe the system, system capacity, and contribution, and the systemic causal role theory will give you whatever answer your intuitions favored from the start. The systemic causal role theory itself does not have the resources to tell us whether any of these statements should be preferred and why. One way to restrict the possible interpretations is to require that system capacities must themselves be functional (this could eliminate frivolous system capacities like “the capacity to smoke cigarettes”). But this would make the account viciously circular.

To avoid the problems traditionally associated with systemic causal role theories, there are two main directions systemic causal role theorists have gone. One is to index function attributions to external perspectives; external perspectivalist theories are considered in Section 3.4. The other is to add the further constraint that the systemic capacity being contributed to must be a form of *self-maintenance*. This type of account is considered next.

[33] And again, Neander’s (1995) solution to indeterminacy problems, described in an earlier footnote, will not suffice because the function attributions are not all connected by interlevel mechanistic “by-relations.”

3.3.3 Self-Maintenance-Based Approach

Moreno et al. offer specific criteria for what types of systems are self-maintaining, and what counts as a contribution to self-maintenance, which are based on their notion of *thermodynamic constraint closure* (Moreno & Mossio, 2015, Chapter 1). To yield an account of proper function, and a distinction between function and malfunction, Moreno et al. must also answer the question, in virtue of what can a part of such a system continue to qualify as belonging to a functional type if that part is no longer contributing to the system's self-maintenance? Mossio et al. claim that "To have functions, self-maintaining systems must belong to a specific class" (2009, p. 825), depending on "the kind of organization they possess" (2009, p. 829). Such organization can persist even when some parts are no longer contributing to the system's self-maintenance; such parts are then said to malfunction. The answer to the Grounding question is then: traits have functions in virtue of the fact that they are part of organized systems that require self-maintenance. The answer to the Demarcation question is: traits are functioning properly when they are actually contributing to the self-maintenance of the larger organized system.

But if the actual presence of self-maintenance is not necessary for individuating such classes of organization, then how are they to be individuated? Almost any type of organized system will be subject to entropic forces and will therefore require maintenance, so there are few constraints on what counts as an organized system. Certainly kinds of organization can be distinguished at varying levels of grain, and since function is relative to organizational class, a generality problem will appear here as well.

Christensen and Bickhard note that "the conditions of self-maintenance vary over an organism's lifetime and across generations (due to environmental variability amongst other things)" (2002, p. 10). Moreno et al.'s way of putting this point is that any given class of organization will be capable of adopting multiple "regimes of self-maintenance"; a regime of

self-maintenance is a “possible specific organization that an individual member of a class can adopt without ceasing to exist or losing its membership of that class” (Mossio et al., 2009, p. 849). But since a regime of self-maintenance essentially amounts to a different *way* of conducting self-maintenance, “functional ascription could vary according to the specific instance of [self-maintenance] that the system is realising at a given moment (what we called a ‘regime of self-maintenance’)” (Moreno & Mossio, 2015, p. 74). This means that functional ascriptions depend not only on the level of generality at which organizational classes are typed, but also the level of generality at which regimes of self-maintenance are typed:

> Each specific regime, orders and levels of closure generate … a distinct set of norms and functions. … [A] given function … could be at work only within a specific regime of maintenance of the considered system, realised, for instance, only in some particular conditions or at a given moment. As a consequence, adequate functional ascriptions should make explicit, in each specific case, which are the regime, order, and level of the closure involved… (Moreno & Mossio, 2015, p. 74)

Further, since the requirements for self-maintenance can change when the environment changes, whether a part is contributing (or can contribute) to self-maintenance may change depending on the organism’s context (Mossio et al., 2009, p. 832). But whether or not the environmental context should be considered to have changed is dependent on the level of generality at which the environment and its features are typed, so a generality problem will also exist at the level of context-dependence of biological function ascriptions.

3.3.4 Lesson: Generality Problems Afflict Non-Perspectival Theories in General

It is not hard to see how other non-perspectival theories of proper function can easily run into such problems as well. For instance, the modal theory of function (Nanay, 2010; cf. Bickhard, 2000, pp. 116–117) defines function and malfunction counterfactually in terms of the closeness of possible worlds. Of course, possible worlds can be typed at varying levels of

grain, and this will drastically impact which ones are counted as being “close” to each other (Enç, 2002, p. 303).

The general lesson of Section 3.3 has been that theories of proper function that assign proper functions to trait tokens, on the basis of how some related non-perspectival token (selection history, system the trait plays a causal role in, self-maintenance regime, etc.) is to be typed, will give rise to generality problems. Without supplying a general principle for choosing the relevant type in given cases, such a theory cannot provide a function assignment with definite content because such assignments will tend to vary widely depending on the generality at which the relevant token is typed. These problems may or may not be insurmountable; but if I am right that such problems do not arise for perspectivalist accounts, they offer an important motivation to give perspectivalism a closer look.

3.4 External Perspectivalism

Though external perspectivalism has not had as much uptake as the non-perspectivalist approaches, defenses of it have begun to appear recently (e.g., Craver, 2001, 2013).[34] According to a simple version of external perspectivalism, the proper function of a trait token consists in whatever an external observer considers it to be. In this bald form, it is obvious how such a view runs into problems: different observers might disagree on what the proper function is, or the same observer might hold differing views on different occasions (Maley & Piccinini, 2017, p. 240). One way around these problems is to relativize proper function to specific observers and specific occasions: the proper function of A for observer B on occasion C is D. But since we do not relativize biological *explanations* to specific observers and specific

[34] An earlier account that is unambiguously perspectivalist was developed by Wimsatt (1972). Because systemic causal role theorists often find it necessary to appeal to things like the interests of scientists to avoid promiscuity objections, they are sometimes read as what I am calling external perspectivalists.

occasions, it is hard to see how this would provide answers to the Demarcation and Grounding questions in a way that accounts for the explanatory role of proper functions.[35]

The usual way of dealing with these problems is to relativize function, not to observers and occasions, but instead to *interests* of external observers (usually scientists). The rationale is that biological explanations are themselves relative to explanatory interests (Machamer, Darden, & Craver, 2000, p. 13), so it should not be problematic if function attributions are relative in this way as well. However, it is hard to see how proper functions can be explanatory if explanatory interests are appealed to in order to give definite content to such functions in the first place; the mere fact that a scientist has taken an interest in something does not impart explanatory power to it.[36] To avoid this problem, the external perspectivalist may respond that explanatory interests themselves are suitably constrained by non-perspectival empirical facts; but in this case, it would seem that such facts are what are actually answering the Demarcation and Grounding questions. Therefore, the external perspectivalist about proper function is trapped in a dilemma: the account either fails to satisfy a key desideratum for an account of function, or it reduces to a non-perspectival theory, in which case generality problems are likely to emerge (Section 3.3).

But there are additional problems for external perspectivalism. Bigelow and Pargetter argue that there is often a mismatch between functions and what we take an interest in:

> it is assumed that biological structures *would* have had the functions they do have even if we had not been here to take an interest in them at all. And some of the effects of structures that we take an interest in have nothing to do with their function. And some functions are of no interest to us at all. (1987, pp. 183–184)[37]

[35] See Massimi (2018a, p. 347) for additional problems for making such a move.
[36] Woodger went further, arguing that attempts to "explain internal teleology by means of external teleology" are no better than vitalism or Cartesian dualism if they leave the external teleology unaccounted for (1929, p. 441; cf. Pattee, 1977, pp. 260–261).
[37] Mark Bedau (1992, p. 37) makes similar arguments against what I am calling external perspectivalism about goal-directedness.

External perspectivalists may yet be able to devise effective responses to such criticisms. My intention in considering the above theories and their weaknesses has not been to make an argument by elimination, but instead to motivate exploration for an alternative theory that does not face generality problems and that does not make proper functions relative to the subjective mental states of external observers. It is to this exploration that I now turn.

3.5 Avoiding Generality Problems: Non-Minded Systems That Have Perspectives

3.5.1 Pattee, Dretske, and Selective Loss of Detail

Generality problems, in general, make it difficult to provide a sufficient answer to what I have called the Grounding question, which, applied to the proper function debate, is: What *makes it the case that* any such mode of functioning should count as a proper function? In order to solve generality problems, any theory that analyzes proper function in terms of some other factors must tell us what it is about any given case that *fixes* the appropriate level of generality at which such factors should be considered.

What kind of thing can “fix” or “pick out” a level of generality in the first place? Clearly, human beings can pick out a level of generality; this is just what Conee and Feldman and Bishop argue is happening when philosophers “apply” process reliabilism to given problems. Arguably, this is because human beings are capable of *abstraction*; they are capable of taking different perspectives on a situation, and characterizing it at varying levels of description, some of which are more fine-grained than others. But most theories of proper function attempt to answer the Grounding question in a way that does not make essential reference to the subjective mental states or cognitive capacities of humans, because otherwise they cannot do

the work referred to in the introduction (naturalizing representation, value, etc.), and they cannot account for the explanatory utility of notions of proper function.

The question then is: what sort of thing is it to *have a perspective* at all, i.e., the sort of perspective that allows us to "fix" *some* level of generality at all? What is special about the human capacity for *abstraction* that is relevant to generality problems? Pattee's answer to this is that *selective loss of detail* is the key:

> All forms of discrimination, measurement, classification, selection, pattern recognition, etc., are accomplished by selective loss of details. Quite generally, it is this selective loss of details from the immense array of sensory inputs and memory stores that makes symbol systems of all types, including all languages, functionally useful. Such discriminations are certainly one of the most primitive and fundamental functions of the brain. (1992, p. 190)

For instance, the ability to discriminate houses from other types of buildings requires a selective loss of detail: it requires sensitivity (or attention) to the details that distinguish houses from non-houses, but insensitivity (or lack of attention) to details that are irrelevant to making such discriminations (e.g., whether the building has four windows or five). By switching to a different perspective, different details are lost or treated as insignificant. The *generality* at which discriminations are made is given by the level of detail that is treated as significant, and the level of detail that is ignored, i.e., *selectively lost*.

But Pattee does not believe that a mind (or even a nervous system) is necessary for this. Relatively simple dynamical systems and devices can exhibit a selective loss of detail in the way that they are sensitive in their interactions with other objects. Similarly, Dretske argues that simple analog/digital devices, by systematically ignoring information at a certain level of detail, are capable of processes that involve assigning tokens to types, such as classification, categorization, or recognition:

> Until information has been lost, or discarded, an information-processing system has failed to treat *different* things as essentially the *same*. It has failed to classify or categorize, failed to generalize, failed to "recognize" the input as being an instance (token) of a more general type. (1981, p. 141)

Dretske argues that it makes sense to speak of such a device as having its own "point of view," according to which multiple tokens are considered to be of "the same" type. Similarly, Popper argued that since no two physical events are ever exactly the same, "for logical reasons, there must always be a point of view … before there can be any repetition; which point of view, consequently, cannot be merely the result of repetition" (2002, p. 59). Dennett (1991) and Ross, Ladyman, and Collier (2007, pp. 220–227) use the word 'perspective' in the same sense as Dretske's 'point of view'. Devices that perform pattern recognition operate with a determinate scheme of sorting tokens into types—a *de facto* perspective on which tokens should get assigned to which types. In fact, in his book-length treatment of the notion of a "point of view," Hautamäki argued that while this notion gets used in many different ways, the commonality is that when a given thing is being considered from a point of view, "certain aspects of [that thing] are considered while others are ignored" (1986, p. 7), i.e., a selective loss of details. On Hautamäki's account of points of view, out of all of the possible "determinables" (i.e., types) that a determinate (token) object or situation may fall under, a point of view represents a selection of certain of these determinables as relevant and others as irrelevant (ibid., p. 65).[38]

Pattee argues that selective loss of details is ubiquitous within biological systems, as it is essential to the working of enzymes; in fact, Pattee argues that ultimately, all

> pattern recognition and selective action [in biological systems] is mediated by enzymes or enzyme-like molecules. This is the case for the cell's sensing of the external environment, for sensing between the cells, and for intracellular recognition of patterns. (1982, p. 172)

[38] My main reason for adopting Dennett's and Ladyman, Ross, and Collier's use of 'perspective' rather than Dretske's, Popper's, and Hautamäki's term 'point of view' here is to highlight that the key advantage that external perspectivalism has over forms of non-perspectivalism can be retained while dropping the dependence on mentality or cognition. Generally, the two terms are treated as synonymous. I would define 'perspective' as an isolated set of pattern-types which the system uses to implicitly carve its own world of interaction (cf. Varela, 1997). My use of 'perspective' is also similar to Devlin's (1991, p. 151) notion of a 'scheme of individuation'. Dennett's (and my) use of 'perspective' should not be confused with his use of 'stance'; a perspective counts as a "stance" for Dennett when it is adopted by minded beings as a "predictive strategy" (1981, p. 15).

Enzymes act as detection devices because of the selectivity with which they bind to other molecules. They have binding sites at which they form electrochemical bonds to specific types of molecules (often other proteins or carbohydrates). The formation of an electrochemical bond is much more like a discrete process than a continuous one; normally a protein either binds to a molecule or it doesn't. Further, enzymes are insensitive to differences between molecules of a given chemical species; either the catalytic reaction is carried out or it is not. Subtle differences in conformation between substrate molecules that bond to the enzyme do not translate to qualitative differences in catalytic reaction; such differences are effectively "ignored" by the enzyme:

> out of all the microscopic collisions with … an enzyme, only very special ones are capable of triggering their catalytic function, and furthermore when this function is triggered only a very limited or simple result takes place. This is in contrast to ordinary dynamical systems where almost any collision results in a complex perturbation spreading through the entire system with no coherent result whatsoever. (Pattee, 1973c, p. 43)

In this sense, protein binding, which is the basis for enzymatic reactions and intracellular signaling, counts as a fixed practice of recognition and response that consistently sorts substrate molecules, as well as operates on them, at a fixed level of generality. For these reasons, Pattee (e.g., 1982, p. 172) argues that enzymes can be said to *recognize* or *classify* molecules; enzymes have their own perspectives on which differences between molecules are significant and which are not (which determinables are relevant and which irrelevant, in Hautamäki's terminology)—perspectives on how to sort molecule tokens into types (viz., substrate and nonsubstrate).

An external perspective can be a way of sorting tokens into types, and so indexing the sorting of tokens into types to such an external perspective screens off alternative interpretations. But if Dretske and Pattee are right, then a theory of proper function need not make essential reference to human minds or sophisticated cognitive systems; it could instead ground the normativity of function in the classificatory or recognitional capacities of certain

kinds of much simpler systems. But even given Pattee's claim that biological systems contain vast numbers of classification and recognition devices, it remains to be seen how these resources can provide a basis for norms of biological performance, i.e., proper function. In the next section, I discuss work by Haugeland that attempts to demonstrate how basic classification and recognition capacities can provide the basis for norms of performance.

3.5.2 Selective Loss of Performance Detail: Haugeland and Censoriousness

In his 1990 paper "The Intentionality All-Stars," Haugeland's goal is to provide an account of the normativity of intentionality. Though he is not providing an account of biological proper function per se, his theory is worth examining in this context because of the more general insights it can offer about what is needed to avoid generality problems about norms of performance.[39]

According to Haugeland, normative standards of performance can spontaneously emerge within a community of "versatile and interactive creatures, not otherwise specified except that they are conformists" (1990, p. 147). Haugeland discusses two forms such conformism might take, i.e., two mechanisms of interaction by which such creatures might participate in and sustain a community of shared performance norms. The first of these is *imitativeness*, i.e., the tendency to imitate the behavior of the other "creatures." Turner (1994) argues that imitativeness alone cannot provide a ground for the sharing of norms. This is because imitating is subject to interpretation: there are many ways to interpret another's behavior, and therefore many different ways to imitate it. Again, a generality problem arises: whether or not A's behavior is an imitation of B's behavior depends on the level of generality or

[39] Haugeland (1990, p. 147) credits Heidegger, Sellars, and Brandom as forerunners of the position he is advancing.

grain at which the behaviors are typed, i.e., at which behaviors are considered as being *the same* for purposes of imitation:

> What is needed to preserve the concept of sameness in the case of open-textured rules is … some way of distinguishing variation of habits, or the acquisition of a variant rule, from the acquisition of the same open-textured rule. (Turner, 1994, p. 76)

In other words, what is needed is a shared scheme for sorting and discriminating the performances of others. Turner and Haugeland both argue that what is needed to fix a norm is not merely a tendency of imitativeness, but a form of what Haugeland refers to as *censoriousness* that fixes and categorizes the relevant behavior *types*. Turner notes that, for this reason, norms "are identifiable only by observing what happens if they are breached" (1994, p. 28). Haugeland says that by censoriousness, he means "a positive tendency to see that one's neighbors do likewise, and to suppress variation" (1990, p. 147). He does not only have negative kinds of censoriousness in mind:

> What in general counts as censorious acceptance and rejection? Whatever community members do that promotes conformism: if they smile at conforming performances, and smiles promote repetition, if they fire electric shocks at those who err, and these shocks discourage aberration, then smiles and shocks are devices of censorship. And who are the members of this community? The members are basically whoever is brought into conformity with the rest of the group, and thereupon participates in the censoring of others. (1990, p. 148)

Haugeland explains that such censoriousness can form the grounding of a socially-based norm if it is based on a fixed practice of recognition and sorting of behavior.

> Imagine, for instance, that the rules of chess were not explicitly codified, but observed only as a body of conformist norms—"how one acts" when one plays chess. So, it is proper (socially acceptable) to move the king in any direction, but only one square at a time. For this to be a norm, players and teacher/censors must be able to "tell" (that is, respond differentially to) which piece is the king, what are the squares, what counts as a move, and so on; thus, the presupposed sorting of circumstances is effectively a sorting of items, features, and events within those circumstances. Meanwhile, according to other norms, the king must be protected when threatened, cannot be exposed to capture, can castle under certain conditions, and so on; and, crucially, for all of these norms, it's the same instituted sort ("king") that's involved. Hence, the norms themselves are interdependent via depending on the same sorting of circumstances (items, features, …). (1990, pp. 151–152)

As long as the sorting and categorizing is done in the same way by participants, there is no generality problem; which piece is counted as a "king" is not left open to interpretation based on differing levels of grain. In order to be playing chess at all, in this community, one must first of all attend to the same level of specification or detail as all the other participants when counting something as a "king" or not. But Haugeland stresses that the categorization must not only happen at the level of recognizing and categorizing items, features, and events; it must also happen at the behavioral level, so that input stimuli are recognized and categorized and then mapped to a categorical response: "norms have a kind of 'if-then' character, connecting sorts of circumstance to sorts of behavior," (1990, p. 151).[40] Haugeland argues that because conformism (in his sense) is the type of thing that serves not only to determine the behavior of individuals in the community but also that sets the framework for *typing* circumstances, behaviors, and conditions that hold between them, as well as for discriminating deviations, conformism can provide the necessary grounding for performance norms:

> When behavioral dispositions aggregate under the force of conformism, it isn't herds that coalesce, but *norms* … distinct, enduring clusters of dispositions in behavioral feasibility space, separated in that space by clear gaps where there are no dispositions … . The community-wide classes of similar dispositions that coalesce under the force of conformism can be called "norms"—and not just collections or kinds—precisely because they themselves set the standard for that very censoriousness by which they are generated and maintained. The censure attendant on deviation automatically gives these standards (the extant classes themselves) a *de facto* normative force. (1990, p. 149)

Haugeland stresses that conformism, on his view, does not require the capacity for following explicit rules, but neither does it consist in mere causal regularity. Here, Haugeland trades on a

[40] Schroeder has argued that

> Regulating involves creating a rule of some sort, as the word 'regulate' suggests. This is the sort of activity that can be expected to create norms, for a rule is just one sort of norm. Hence it should be no surprise that a regulated object is subject to a norm of performance: that it has a function. (2004a, p. 118)

However, as will be discussed in Section 4.4.1, not all forms of regulation rely on fixed typing schemes that define *sorts* in Haugeland's sense. Simple forms of negative feedback, for example, are insufficient for normativity because it is a matter of interpretation how far the variable being controlled has to be from the set point until malfunction should be ascribed.

distinction that was earlier mapped out by Sellars, who distinguished between "merely conforming to rules," "obeying rules," and a third category situated in between the first two that Sellars refers to as "pattern-governed behavior" (1954, pp. 326–327). Haugeland also argues that conformist norms should not be confused with conventional norms; on his view, the latter require mental states:

> The difference between norms and conventions lies in this explanatory appeal: conformism does not presuppose any prior beliefs or preferences on the part of individual conformists, and hence the persistence of norms cannot be explained in terms of agents' interest maximization or rational choice. Indeed, norms need not be, or even seem to be, in any way beneficial either to the individuals or to the group; the mechanism of conformism is completely blind to the character or merits of the norms engendered.[41] (1990, p. 150)

For these reasons, the information processing capacity of "creatures" constituting the conformist community can be extremely thin. They do not have to be human beings: they could be very simple organisms or machines—anything that has the capability to perform recognition and enact a rule-like (pattern-governed) mapping between recognized circumstances and behavioral response categories.

3.5.3 Division of Labor

Whereas Haugeland's account depended on a community of homogenous "creatures" that are all monitoring and censuring each others' behavior, the entities doing the censuring (call these the "alpha creatures") might be distinct from those which are censured ("beta creatures"). Consider an example in which all of the creatures are robots in a fully automated factory (perhaps they were all designed by intelligent beings, or suppose that their existence and placement within the factory is a chance occurrence, like the parts of Dretske's "Twin Tercel," 1995, pp. 141ff). The beta robots are the workers. They perform the work in the

[41] Stated another way: Conformist norms need not be *categorically binding* (in the sense of Copp, 2015), and need not be constrained by categorically binding norms.

factory, assembling and transporting components, and they each have a task that they usually perform successfully. But they sometimes develop problems causing them to fail at their tasks. For example, their internal components may become corroded or damaged, there may be lurking bugs in their software, or their parts may need recalibrating. Fortunately, the alpha robots are there to roam the factory and monitor the performance of the beta robots. When the beta robots are doing their jobs properly, the alphas simply continue vigilantly roaming through the factory. When they detect that a beta robot is not performing its job properly, the beta robot is censured. This may take many forms: perhaps taking it offline and sending it to a repair bay, or taking it offline and sending it to a scrapping or recycling area, or even sending a signal to the beta robot that stimulates a self-repair activity. Regardless, even in this simple situation, the automated factory may be said to realize a communal regime of "conformist norms" in Haugeland's sense, even if the information processing capabilities of the robots are very simple. Again, Haugeland stresses that in the minimal case,

> conformism does not presuppose thought, reasoning, language, or any other cognitive faculty; the creatures do not in the first instance conform or censure wittingly or because they want to (except "tacitly")—they are simply built that way. (1990, p. 148)

Nonetheless, by enacting a specific framework for typing circumstances, behaviors, and conditions, as well as what count as deviations, and censuring these deviations, the censuring patterns of the alphas set a normative standard for the factory as a whole. Given the background of such norms of performance that are put into practice by the alphas, the beta robots in the factory may be said to have proper functions; the standard for whether a given beta robot is functioning properly is set by the censuring patterns of the alphas.

Note that by putting a certain set of standards into practice in the factory, the alphas may be said to collectively *have a perspective*: their pattern of censuring constitutes a tacit perspective on which differences in performance are significant and which ones insignificant. An external observer may take a different perspective and decide that the way the alphas work

is not *objectively* best for the factory. The perspective possessed by the alphas may not be "objective" in this sense[42] (as we saw, due to generality problems, there probably cannot be a theory of what counts as proper function in this non-perspectival sense), but it is objective in the sense that it is independent of any *external* observer. The system in question, by itself, realizes a normative standard. The collective activity of alpha robots also provides an objective criterion for delineating what is part of the system and what is not.

Such norms of proper function are not etiological or history-dependent, and they are not dependent on the perspective(s) taken by external observers. They depend on the level of generality at which certain circumstances, behaviors, and conditions are typed, but instead of "starting with the intuitively correct judgment about a [trait's proper function] and then cherry picking an account of [the organism's selective regime, selection history, possible worlds, regime of self-maintenance, the functional characteristics of the larger system, hierarchical levels, etc.] that yields that judgment" (again paraphrasing Bishop), it is the alpha robots themselves that *fix* the relevant level of generality. For this reason, by grounding proper function in the censuring activities of alpha robots, the generality problems can be avoided. In the next section, I discuss how the theory applies to real biological systems.

3.6 Performance-Monitoring and Censuring Mechanisms in Subcellular Biology

The cells of biological organisms, including even single-celled organisms like bacteria, contain highly sophisticated molecular machinery. The everyday functioning of any given cell involves many different kinds of chemical reactions that are tightly orchestrated and coordinated in ways that biologists are only beginning to understand. There are many ways

[42] Or, again in Copp's (2015) terminology, the performance norms grounded by the censuring patterns of the alphas are not categorically binding.

that such orchestration and coordination can (and does) fail, and like an automated factory, cells are highly dependent on certain components (proteins) that are capable of sensing whether certain other components are doing their job properly. As discussed in Section 3.5.1, the enzymatic binding and catalysis by which proteins perform their work also counts as a fixed practice of recognition and categorical response. Just as chess players consistently categorize chess pieces at a fixed level of generality, enzymes sort substrate molecules at a similarly fixed level of generality. These quality control proteins can trigger various types of responses, ranging from throttling down upstream processes to reduce the malfunctioning component's workload, to triggering a process called *programmed cell death*, in which the cell itself commits suicide in an orderly fashion to prevent the malfunctioning cell from causing further problems for the larger organism.

One example of this is the *unfolded protein response* that is triggered by several different types of "performance monitoring" proteins in the endoplasmic reticulum (ER). An important task that many cells perform is to synthesize proteins that will be excreted from the cell to perform functions elsewhere. These proteins (among others) are synthesized in ribosomes that are embedded within the ER. Protein synthesis involves stringing together component amino acid molecules, and requires the string of amino acids to be folded into the right shape (important because the capacity of proteins to do work is largely a matter of how they are shaped). The folding process can occur spontaneously during synthesis, but is sometimes also assisted by helper proteins called *chaperones*. Problems may occur during the folding process, whereby a non-folded or misfolded protein results, or a properly folded protein can also become unfolded or misfolded as a result of various conditions in the cell, like excessive heat or pH levels (the unfolding of a protein is sometimes called *denaturing*). There are several types of proteins in the ER that are able to sense unfolded or misfolded proteins, including Ire1, ATF6, and PERK (Galluzzi, Bravo-San Pedro, & Kroemer, 2014).

When too many unfolded proteins are detected by such “sensor” proteins, several types of responses may be triggered. A signal may be sent to the nucleus of the cell causing it to produce more of the chaperone molecules or to enlarge the ER to increase its processing capacity. A signal may also be sent that causes the upstream processes of protein synthesis to be slowed down (to give the ER a chance to play catch-up, and hopefully reduce the number of unfolded or misfolded proteins). These signals may also halt the progression of the cell cycle, preventing the cell from making further progress towards mitosis, or cell division (since if this cell is malfunctioning, then the resulting cells may not be properly formed). When conditions are bad enough, these sensors can even send a signal that triggers *mitochondrial membrane permeabilization* (MMP), in which “executioner” enzymes are released from the outer membrane of the mitochondria. These executioner enzymes, called caspases, are like a controlled demolition team that systematically degrades the components of the cell, causing it to self-terminate without provoking an immune response or causing problems in the surrounding tissues (Hotchkiss, Strasser, McDunn, & Swanson, 2009).

Other examples of malfunction detection and response have been found in other organelles, including the mitochondria, nucleus, Golgi apparatus, and lysosomes (Galluzzi, Bravo-San Pedro, & Kroemer, 2014). For example, the SOK1 sensor protein in the Golgi apparatus detects malfunction by responding to elevated levels of *reactive oxygen species* (ROS), which are a potentially dangerous byproduct of certain intracellular reactions (Nogueira et al., 2008). Instead of triggering programmed cell death, another way that cells respond to malfunction is by causing the orderly disposal of malfunctioning organelles in a process called *autophagy* (Okamoto, Kondo-Okamoto, & Ohsumi, 2009; Youle & Narendra, 2011).

Even when the damage repair mechanism is not known, its existence can be ascertained. For several decades, biologists have known that when filaments are sheared off of flagella, bacteria are able to replace them. Kelly T. Hughes and colleagues found that flagella can regenerate from the same hook/basal body of a flagellum that was sheared off. Through

clever experiments they determined that a separate mechanism for *repairing* malfunctioning flagella must exist, rather than simply retriggering the same mechanism that initially builds the flagellum, although this separate mechanism is not yet known (Rosu & Hughes, 2006).

These examples have several features in common. In each case, there is some situation indicative of performance which is either registered as present or not present (e.g., malfunctioning flagellum, too many unfolded/misfolded proteins, too many ROS molecules being produced/too much oxidative damage, malfunctioning or superfluous mitochondria), triggering an unequivocal all-or-nothing type of response (e.g., flagellar filament replacement, autophagy, halting the cell cycle, apoptosis). In other words, the recognition and typing of a performance token, an implicit if-then mapping from this recognized performance type to a categorical censuring response, and the context-dependent execution of that response, all the ingredients needed for a community of "creatures" to realize norms of performance based on censorious acceptance and rejection. Since what is being categorized is the performance of some subsystem or component, each of these processes can ground a distinction between functioning and malfunctioning for that subsystem/component.

3.7 Macro-Level Functions: Composite Recognition and Response

In the previous section, the focus was on individual protein molecules that act as "alpha creatures" at the subcellular level. But it might seem that there is no equivalent at the macroscopic level that can ground macro-level proper functions, such as the mammalian heart's function of pumping blood. After all, there is no individual system in mammals we can

point to that that acts as an alpha creature, measuring overall performance of the heart and "censuring" the heart when it malfunctions.[43]

Recall from Section 3.5.1 that what is special about enzymes that allows them to act as alpha creatures is their selective sensitivity to *types* (chemical species) at a specific level of detail: differences in conformation between substrates of the right chemical species for that enzyme are ignored, so that enzymatic reactions implement the kind of fixed practice of recognition (in the Dretskean sense) and categorical response that can ground a normative perspective. While such enzymatic activity is ubiquitous in biological systems (Section 3.6), similar detection and censuring systems above the cellular level may not be nearly as common. However, I do not believe that this shows that most of the macro-level functions we attribute to biological traits do not exist. In this section I argue that instead of arising from the activity of individual, macro-scale alpha creatures, most macro-level proper functions emerge from the collective activity of micro-level alpha creatures.

To understand how macro-level proper functions (e.g., of an organ like the heart) can arise from the activity of the kinds of micro-level alpha creatures described in the previous section, begin again with a simple example. Suppose that there is a device that detects whether there is a '1' in a certain position, and another device that detects whether there is a '5' in the position next to it. Each device responds categorically in one of two ways depending on whether the condition in question is met. But when you consider both devices together, there are four conditions that the devices are *collectively*—not singly—responsive to: present 1 and absent 5, present 1 and present 5, absent 1 and present 5, and absent 1 and absent 5. There are similarly four *collective* response behaviors, one for each of these conditions. In this way,

[43] The hypothalamus plays a key role in regulating heart rate, but this is not quite the same thing: the hypothalamus only makes adjustments to normal heart functioning. It has no way to detect and "censure" the heart based on macro-level conditions of heart malfunction, like congestive heart failure or cardiac arrest.

we can instead look at the two detection devices as comprising a single, more complex detection device.[44]

Now suppose that the absence of a 1, in the case of the first detector, or the absence of a 5, in the case of the second detector, results in negative censuring behaviors (perhaps they are monitoring and censuring the mechanisms responsible for producing the 1s and 5s). The detectors are then each a type of alpha creature. If so, then it equally makes sense to consider the combination of the two detectors as a single, more complex alpha creature that monitors whether the sequence (1, 5) is present. It is monitoring the larger system that includes the subsystems that produce the 1s and 5s, and enforcing a standard in which the production of (1, 5) sequences constitutes proper functioning, and other conditions constitute varying degrees of malfunction.

The larger point here is that the activity of many alpha creatures can collectively realize a higher-level, and more complex, standard of functioning—not due to the fact that the higher-level censuring is causally dependent on the low-level censuring, but because it is *constituted* by it. The higher-level response is categorical and involves a selective loss of detail precisely because the lower-level responses are categorical and involve selective losses of detail. In fact, the detail that is lost at the higher level *just is* the sum of the detail that is lost at the lower level. As long as the lower-level alpha creatures operate with a fixed level of generality, any normative censuring standard that is additively built up from them will also operate with a fixed level of generality. Ultimately, activity that passes muster for many different types of alpha creatures at the cellular level *just is* activity at the macro scale when you add it all together. The alpha creatures collectively define a standard of what passes muster at the macro-level, and therefore define what proper functioning consists in (when no censuring activity is triggered) or what varying degrees of malfunction consist in (when varying amounts of negative censuring

[44] In fact, this is the basis of parallel distributed (Selfridge & Neisser, 1960) and agent-based (Rosin & Rana, 2004) pattern detection techniques.

activity is triggered). Just as in the simple initial example, the function of the combined subsystems was to produce sequences of (1, 5), the function of the heart is to pump blood because that is the collective activity that the detectors are collectively responsive to. The detectors, considered collectively, realize a perspective on what counts as proper functioning for the heart as a whole. In fact, this is the reason why we don't find a single, non-composite macro-level alpha creature that monitors the heart as a whole: such a macro-level alpha creature is not necessary and would be redundant, because that role is already fulfilled by the collective, finely orchestrated and emergent self-organizing activity of lower-level alpha creatures.

Go back to the automated factory example, and suppose it is a car factory. It might be the case that there isn't any particular alpha robot whose job it is to monitor the factory as a whole and detect whether it is producing entire cars properly or not. This might be because the operations of the individual alpha robots are so well orchestrated (without there needing to be a non-composite "orchestrator") that their activity collectively adds up to the implementation of a standard-verifying mechanism for the whole factory that works as well as a single, holistically functioning alpha robot would. In this way, an organ like the heart is like the car factory.[45]

Space does not here allow a detailed critical analysis of this way of grounding macro-level proper functions. However, since on this account, proper function is always derived from the detection of individual enzymatic detectors (and other Dretske-style recognition devices that might exist in biological systems), an objection might run as follows: Even at the level of

[45] Unfortunately it isn't possible to go into empirically detailed concrete biological examples of this, simply because we don't know in detail about how the activities of thousands of "alpha creature" individual proteins collectively sustain the organization of a system like the heart. It would require reverse-engineering the heart in *total* molecular detail, all the different genes that get expressed under what circumstances (and exactly what molecular pathways they are triggered by) in all the different types of cells of the heart and what their intercellular and intracellular effects are, and all of the non-linear interactions of those effects, etc., something we cannot yet approach. I don't, however, believe that such an exhaustive level of knowledge is necessary for scientists to competently discover and investigate macro-level proper functions, just as knowledge of molecular structure is not necessary for competently discovering and classifying crystals (Polányi, 1958, pp. 43–48).

subcellular detection enzymes, it is a matter of interpretation whether (in the case of the enzyme that detects mitochondrial damage) increased ROS is being detected, or whether incomplete oxidative phosphorylation is being detected. This is similar to the argument that there is no determinate representational content in frog vision when a frog is detecting a fly, because there are equally good reasons to think that what is being registered by the frog's visual system is the presence of 'food' or the presence of 'small dark moving things' (Goode & Griffiths, 1995, p. 100).

However, this objection conflates two different types of problems. One pertains to the question of whether there exists a determinate category *for the frog* (i.e., by the frog's "lights," or according to its own perspective) of what is being detected. The other pertains to the question of how to translate such a category into the ordinary language of humans.[46] The latter problem is an important one to solve if one wishes to ground the full richness of representational content of human cognition in basic perceptual detection capabilities, which is certainly not being attempted in this chapter. There may very well be no fact of the matter whether we should characterize what the frog is detecting as 'food', 'flies', or 'small dark moving things'. But it may be true at the same time that since the frog has no cognitive means of distinguishing between these situations, and only has a single category corresponding to its more basic recognition capacity, no corresponding ambiguity exists from the point of view of its own content-assignation scheme (cf. Dennett, 1996, pp. 42–43). The categorical discrimination necessary for such ambiguity to arise outstrips the frog's own classificatory repertoire.

Similarly, even if we can conceptualize the performance-monitoring enzyme's simple classificatory repertoire in different ways, this does not mean there will be a corresponding indeterminacy in the proper function grounded by the enzyme's censuring activity. The level of

[46] For a thorough discussion of this type of conflation, see Evans (1975).

generality that the enzyme works with is fixed, and is simply more coarse-grained than the conceptual discriminations that we ordinary language users have at our disposal.

3.8 Conclusion

Because performance-monitoring subsystems that recognize and respond to malfunctions are widespread in biological systems, all biological systems operate with their own implicit perspectives on what constitutes proper functioning. These implicit, operational perspectives can provide definite content to a theory of proper function because they amount to a fixed scheme for sorting performance tokens into types. The *proper function* of a given biological subsystem consists in the activities undertaken by that subsystem that are enforced by performance monitoring detection and response systems. This is an internal perspectivalist account of proper function: what counts as the proper function of a biological subsystem or trait is a matter of the perspective that the biological system, itself, has on what counts as proper functioning for that subsystem.

One key problem faced by each of the major non-perspectivalist positions in the biological function debate is that they assign functions to traits based on the type that some related token (system, selection history, self-maintenance regime, etc.) falls under, but the token in question falls under many such types and this gives rise to conflicting function assignments. As we saw, this is very similar to a problem that has been faced by process reliabilism. What was needed was a theory of proper functions that includes sufficient resources to say what determines the relevant level of type generality in given cases. Internal perspectivalism is the theory of biological function that can meet this key desideratum. The question of when a mechanism is functioning or malfunctioning (the Demarcation question) is one that can be answered by looking at that mechanism in the larger context of the organism and its own quality control “alpha creatures,” which are ubiquitous in biological systems. The

larger system that includes the mechanism in question and the quality control “alpha creatures” can ground this type of normativity (the Grounding question) since their recognition and response patterns will operate with a specific, implicit categorization scheme that isolates a specific level of generality. The present account therefore provides the way to naturalize proper function without reference to natural selection, natural design, or fitness, and without resorting to external perspectivalism.

Acknowledgement

Chapter 3, in full, is currently under review for publication as a journal article under the title “Internal Perspectivalism: The Solution to Generality Problems about Proper Function and Natural Norms.” The dissertation author was the sole author of this paper.

Chapter 4
What is Control?
An Internal Perspectivalist Account

4.1 Introduction

An enormous mismatch exists between the quantity and variety of ways that the notion of *control* is appealed to in philosophy and other fields, on the one hand, and the quantity of attempts to analyze the notion of control, itself, on the other. Consider a few examples of the former:

- Rescher (1970, p. 248) among others has argued that control is what makes the difference between action and mere behavior.
- Frankfurt (1978) argued that online control of behavior is what determines whether an action is performed freely.
- Fischer (1994) argued that control is central to the distinction between persons and non-persons.
- Fischer (Fischer & Ravizza, 1998) has made guidance control the centerpiece of his influential compatibilist account of moral responsibility.
- Bishop (1989) and Schlosser (2007) attribute the main problem for causal theories of action, namely the problem of deviant causal chains, to the fact that such theories do not properly account for an essential ingredient that any theory of action must possess: agential *control*.
- Cybernetics and control theory continue to provide insight and theoretical tools for understanding

    - perception (Powers, 1973),
    - attention (Carver & Scheier, 1981),
    - motivation (Toates, 1975),
    - representation (Sloman, 1996),
    - cognition (Dretske, 1986; Hooker, Penfold, & Evans, 1992),
    - mental content (Grush, 2004),
    - the neuroscience of perception and action (Arbib, 1981),
    - intelligence and thought (in the context of robotics; Brooks, 1991), and even
    - consciousness (Sayre, 1976).
- Sterelny (2001) flatly asserts that “Minds are control systems” (cf. Sloman, 1993).
- The notion of control lies at the center of Morris’s (1946) influential theory of signs and symbols.
- Robertson and Powers (1990) argue that control theory can serve as the basis for an overarching psychological theory.
- In a widely used AI textbook, Russell and Norvig go as far as to say that “the concept of a controller in control theory is identical to that of an agent in AI” (2010, p. 59).
- Redgrave, Prescott, and Gurney boldly claim that “animals can be viewed as control systems” (1999, p. 1011).
- Even more broadly, Pattee has argued in many of his works (e.g., 1973b) that control is what differentiates living from non-living systems.

Stear writes:

> That control paradigms are useful in describing and characterizing a great many of the processes and functions occurring in living systems *in qualitative terms* is

> now well established. … This widespread qualitative use of control concepts to characterize the behavior of living systems illustrates the general pervasiveness and universality of control concepts—whether they are applied to living systems; to the control of nonliving systems such as chemical plants, machines, robots, and so on; or to mixed systems, such as man-machine systems in which humans control the machines. (1987, p. 352)

Despite the obvious importance of control for such a diverse range of topics, there have been very few attempts to understand control itself. There are many discussions that try to analyze what kind or amount of control is necessary to solve specific kinds of problems, or how to implement such solutions. These fall into three broad categories:

- Control in artificial systems (this includes most of what is known as "control theory"); e.g., Mackenroth (2004); Franklin, Powell, and Emami-Naeini (2018)
- Control in biological systems in general; e.g., Milsum (1966), Purich (2010), Khoo (2018)
- Cognitive, agential, and/or intentional control; e.g., Mars, Sallet, Rushworth, and Yeung (2011); Fridland (forthcoming)

But there have been few attempts to understand what all of these domains have in common, or what it is about control that makes it such a widely applicable concept. As a result, there is little crosstalk between the domains, and discussions that attempt to bridge between the domains are often unsuccessful due to the lack of a consistent and rigorous conceptual scaffolding.

What this chapter offers is an analysis of the concept of *control* that aims at being truly domain-general. On the present account:

> A **controller** puts into practice a determinate perspective on what kind of influence (what changes should be made to which control parameters) is appropriate under what conditions. **Control** is influencing (or being disposed to influence) the right aspects of something (i.e., that which is "being controlled") in the right ways to the right extent by the right means at the right times, where what is 'right' in a given context may be determined either by an external

perspective or by the system's own perspective (here using 'perspective' in the sense explained in the previous chapter).

"Influence" will be understood as the exercise of directional causal powers as defined in Chapter 2. I begin in Section 4.2 with an introduction to the notion of an *observer-worker system*, which is a generalization of the notion of an "alpha creature" introduced in the previous chapter. I then, in Sections 4.3 and 4.4, introduce a distinction between autonomous and non-autonomous controllers. Whereas the status of non-autonomous controllers as being capable of control is dependent on an external perspective, autonomous controllers are not dependent on external perspectives in this way. The set of systems that are autonomous controllers is a subset of those systems that are *observer-worker systems*. To be as general as possible, *control* is then defined only in terms of basic concepts introduced in Chapters 2 and 3 as well as terminology from control theory. In particular, it is the presence of a *control parameter*, that is, some common dimension of variability that the output behaviors influence in different ways, that separates controllers from other types of observer-worker systems that are not controllers.

After that, in Section 4.5 I discuss several characteristics associated with controllers: degrees of effectiveness, degrees of control, and degrees of malfunction. It is important to gain a clear understanding of these notions because confusions between them often create difficulties for discussing control across disciplines or domains. In Section 4.6, I present a partial taxonomy of kinds of control systems and discuss how they are related. Importantly, in that section I introduce a distinction between *variable*, *metamorphic*, and *composite* controllers. These are three very different kinds of complexity that control systems can have; this distinction will be important for understanding how agential controllers (i.e., *agents*) are different from other types of complex controllers, as well as understanding the differences between *kinds* of agents (this will be the topic of the remaining chapters of the dissertation).

Finally, in Section 4.7, I briefly discuss several other attempts to define the notion of control and how they compare from the present account.

4.2 Observer-Worker Systems

In this chapter I will argue that the alpha creatures of Chapter 3 and control systems are both special cases of something more general: a *worker*. Physicists have two distinct notions of work: the kinematic notion and the thermodynamical notion. On the kinematic understanding, work occurs anytime there is a force acting through a distance; the "work" involved is simply the quantity of energy transferred. Anything that can generate force (e.g., by possessing mass) counts as a "worker" in this sense. I am instead interested in the thermodynamical notion. Thermodynamics deals with concepts like *heat*, *coherent* motion, and *entropy*. Heat is disorderly or "incoherent" particle motion; the mere dissipation of heat, by itself, is not considered to be work in the thermodynamical sense (Atkins, 1984, p. 198). The clearest examples of *work* in the thermodynamical sense are ones in which energy is constrained to flow in a way that is orderly and useful (Salthe, 2007).

How orderly does a release of energy have to be in order to count as work? For what purpose, or for whom, does it need to be useful? There is no universally applicable standard.

Sometimes thermodynamical work is defined in terms of a distinction between microscopic and macroscopic forces. One recent thermodynamics textbook defines 'work' as follows:

> Work is a process of transferring energy to or from a system in ways that can be described by macroscopic mechanical forces exerted by factors in the surroundings, outside the system. Examples are an externally driven shaft agitating a stirrer within the system, or an externally imposed electric field that polarizes the material of the system, or a piston that compresses it. (Shah, 2018, p. 18)

But again, where to draw the line between "micro" and "macro" will depend on the context. There is no universally applicable standard for how this should be done; nature has no such joint at which to carve. Further, in biological systems, work is often done by individual protein molecules, which are microscopic.

Normally, assumptions about which energy flows will count as "work" are built into the way a thermodynamical problem is defined; the person defining the problem will have purposes or standards in mind. These facts might lead one to adopt an external perspectivalist position about thermodynamical work: something's counting as a worker depends on whether an external observer takes a certain perspective.

This leads to familiar difficulties raised in Section 3.4. What about systems that no one takes an interest in? What if the universe did not contain any cognitive agents capable of forming the thermodynamical concept of work, and therefore did not contain anyone that could provide the needed external perspective for something in that universe to count as work? Would thermodynamics simply not exist in such a universe? This seems tantamount to saying that *order* would not exist in such a universe; but this seems false. Perhaps it's not that *order* wouldn't exist, but instead that there would not be anything that could define a line between what should count as "ordered" and what should count as "disordered."

In fact, thermodynamics defines a notion of *entropy* that does not rely on any such dividing line. Boltzmann defined entropy by the following equation:

$$S = k \ln W$$

where S is the entropy, k is a constant (known as Boltzmann's constant), and W is the number of equiprobable configurations that the system can take on (Atkins, 1984, Chapter 4). This equation can tell us the *degree* of order (or of entropy) that a system possesses, but it does not provide a *standard* for distinguishing instances of heat transfer from instances of work.[47]

[47] In any case of heat dissipation within a closed system, there is an overall increase of entropy, but there can also be *local decreases* of entropy within (thermodynamically open) regions of the system. However, not every local decrease of entropy counts as work. As a large system of particles (for example, water in a kettle cooling down after it has been removed from a heat source) goes to equilibrium, it will not do so in a perfectly uniform way. There will be many local chaotic up and down fluctuations in temperature (Atkins, 1984, p. 82), but these minor fluctuations away from local equilibrium are generally not counted as cases of temporary, local work. This is why thermodynamics texts usually characterize work as "coherent" motion (e.g., Atkins, 1984, p. 48).

What is needed for an account of work is a solution that is analogous to the one offered in Sections 3.5–3.6. A system can ground its own simple standard of proper function by means of its own capacity to recognize and differentially respond to the performance tokens of the beta creatures. But in the present case, we need to ground an even more generic standard: a differentiation between work, on the one hand, and all non-work releases of energy, on the other. Consider again the alpha creatures of Section 3.5. What amounts to thermodynamical work, according to its perspective? Arguably, its organization and operation implicitly define at least three categories of work:

- Properly functioning performances of beta creatures
- Malfunctioning performances of beta creatures[48]
- Censoring behaviors

The performances of beta creatures count as work because they are orderly, constrained releases of energy, arranged into patterns that the alpha creature can recognize. The recognitional capacity grounds a perspective on which such patterns count as *coherent* for the alpha creature. Such patterns are treated not as random noise in the environment but as significant patterns that merit a response. But by being constrained in such a way as to implement such behavioral responses, the alpha creature also itself generates thermodynamical work. Such releases of energy also have significance grounded in the organization of the alpha creature itself by *being* (from the alpha creature's perspective) the appropriate response to the detected condition.

[48] It might be thought that a malfunction need not be a case of thermodynamical work; a malfunction could take the form of an undesired dissipation of heat or perhaps the absence of activity. But at minimum, in order for the condition to be recognized *as a malfunction*, it has to be recognized as a result of the activity of the beta creature in question. The beta creature must be recognized as being present, and as being in *some* behavioral state. Whatever form this takes, it will necessarily need to consist of something other than random particle motion, i.e., a coherent configuration of matter and energy that persists (so as to be detectable) over some period of time.

Note that even though the beta creature's performances count as work, they do not (necessarily) count as work solely in virtue of the beta creature itself; it is the perspective of the alpha creature that grounds the operative distinction of work/non-work that applies to the beta creature.[49] The important lesson here is that in order for a system to be capable of grounding a distinction between work/non-work, it has to be capable of (a) *recognizing* instances of work, and (b) *producing* instances of work. In other words, it has to *be* an observer (in the thermodynamical sense) and it also has to *be* a worker (in the thermodynamical sense).[50] I will refer to such systems as *observer-worker systems*.

But the class of observer-worker systems includes more than just alpha creatures. It includes any system that is capable of recognizing (in Dretske's and Pattee's sense, involving selective loss of detail) patterns as falling under a certain type, and producing a categorical behavioral response when and only when that pattern type is recognized. As was the case with alpha creatures, such a system will have a repertoire of patterns it can recognize (a repertoire containing only a single pattern in the simplest case) and a repertoire of response behaviors (again, that might consist of only one behavior in the simplest case).[51] It will be selectively insensitive to those details that do not pertain to whether the pattern type in question is present or not. The details of motion that it is insensitive to will count as *incoherent* motion for the observer-worker.

By virtue of the fact that something *is* an observer-worker system, then, there will be a corresponding perspective, an ontological carving-up of the universe, had by that system. This ontology will include a range of potential environmental states that can be sensed (*situations*).

---

[49] Exceptions to this will occur when the beta creature is itself an *observer-worker system*, a concept to be defined shortly.

[50] A standard of what counts as an observer therefore requires or implies a standard of what counts as work, and vice versa. This fact was realized by Popper: "Observation is always selective. It needs a chosen object, a definite task, an interest, a point of view, a problem" (2002, p. 61).

[51] It might be said that any observer-worker system has its own determinate *umwelt* (in von Uexküll's terminology) or "functional ontology" (in the terminology of Gallese & Metzinger, 2003).

By doing work (i.e., imposing an orderly pattern onto a flow of energy), they must work with an output categorization scheme. Finally, they must incorporate an internal mapping from input category to output category.[52]

The fact that some amount of detail must be lost for such recognition to occur means that no observer can observe its subject in unlimited detail. This fact is well-known in physics. Since measurement involves making discriminations according to a set standard, measurement counts as a form of observation in this sense.[53] Without a discrete, determinate scheme for sorting quantities into categories (e.g., scale quantities), there is no determinate fact of the matter in any given instance what measurement is being made. Pattee explains the sense in which measurement, like recognition, involves a selective loss of detail:

> Measurement is a very restricted form of perception. To measure something means that you are not measuring everything. More formally, a measurement is a mapping from a physical system to a symbol: but the essence of this mapping is the high selectivity or simplification of the system to only the attribute we have chosen to observe. The problem is this: Can we understand the measurement by decomposing the process in detail? To understand *in detail* would put back into the measuring device all the complexity of interaction that the *function* of measurement requires that the device ignores. In other words, a detailed physical description of a measurement process will look just like any physical interaction of two systems. (1982, p. 174)

In Section 3.5.2, I explained the sense in which an alpha creature works with at least two different categorization schemes: it categorizes the performances of beta creatures into types, and then maps this to a response category (either leave the beta creature alone, or censure it in some way). Haugeland notes that this is true of norms in general: “norms have a

[52] Perhaps the simplest form that such a mapping might take is something like an implicit production rule (see Haugeland, 1985, p. 261, note 15) directly mapping conditions to basic behaviors. Drescher (1991, p. 179) refers to two different basic types of production rule that might be implemented in a simple action selection system: situation-action rules, and situation-result rules. What I have in mind as the simplest type of action selection function would be something like an implicit situation-action rule. See also Minsky’s (2006, p. 20) notion of a “Rule-Based Reaction-Machine.”

[53] Measurement is a concept that has been defined in various ways by different authors (Kuhn, 2009; Tal, 2015; Bradburn, Cartwright, & Fuller, 2017). In this discussion I am following Pattee’s usage, but some authors might argue that a measurement device requires not just any observer-worker system, but one that implements autonomous *coordination* between a dimension of variation and values on a scale; see Section 4.6.2.

kind of 'if-then' character, connecting sorts of circumstance to sorts of behavior," (1990, p. 151). By connecting sorts of circumstance (i.e., potential measurements) to sorts of behavior (potential orderly patterns that can be imposed on the environment in response to the measurements, i.e., work), the observer-worker system is not only the minimal physical system capable of grounding the notions of observation, measurement, and work, it is also the minimal system that is intrinsically normative, i.e., that grounds its own normative standard. The alpha creature of Chapter 3 is a particular type of observer-worker system: one in which what is being measured are *performances* (acceptableness of performance is treated as a binary variable), and in which such measurements are mapped to two behavioral categories: censure or don't censure. The alpha creature grounds a particular type of normative standard: proper function. Not all observer-workers ground a standard of proper function, but by mapping discrete input categories to discrete response categories, they all ground some normative standard.

A minimal observer inside a universe, then, which I have called an "observer-worker system," minimally requires attunement to and discrete sorting of local environmental variations that is internally and systematically linked to the system's behavioral capacities, which perform work on local environmental conditions. This is very close to what Pattee (1991) calls a measurement-control system, with two caveats.[54] First, Pattee believed that a system had to be *living* in order to be a measurement-control system, and second, he considered such systems to be capable of *control* (i.e., to count as control systems). In contrast, I consider neither of these to be necessary conditions for an observer-worker system. Although in this chapter I will argue that controllers (at least, those which I will refer to as "autonomous" controllers) are, like alpha creatures, a special type of observer-worker system, not all cases of

[54] The notion of an "observer-worker system" being developed here is also in many ways similar to Holland's (2012) notion of a *situated signal-processing agent*, where his notion of an *agent* is (in his terminology) a *classifier system* enclosed within a *boundary*. For Holland, a *classifier system* is basically a system that transforms detected signals into effector outputs. Holland's account is less general than the one being developed here, however, since he is only concerned with systems (which he calls *complex adaptive systems*) composed of multiple agents that produce adaptive behavior in the system as a whole.

measurement and work count as control, because control is a specific type of normative standard that is not grounded by all observer-worker systems. In the next section, I discuss a key account of the nature of control, that of Nicholas Rescher, that will be very helpful for explaining what must be added to the minimal observer-worker system to make it an autonomous controller, and what makes control a special type of normative standard.

4.3 The Normativity of Control: Rescher on Control versus Influence

In his important but underappreciated 1969 article on the nature of control in general, Rescher distinguishes between full control, partial control, and influence. He first divides full control into positive control and negative control: “positive control involves the power to assure a desired result, negative control the power to prevent an undesired result” (1969, p. 331). In other words, if something is not in a position to be able to assure a desired result, or prevent an undesired result, then it is not in full control with respect to that desired/undesired result.

Partial control is a case in which full control is shared between two or more controllers. Consider the apparatus depicted in Figure 4.1. Person A, situated at position 1, and person B, situated at position 2, each have full negative control over the presence of outflow through the pipe: person A and person B can each, individually, prevent fluid from flowing through the pipe and out of the reservoir. But each person only has partial positive control over the flow. No one person can assure the flow of fluid out of the reservoir and through the pipe, but they can do this jointly, so they jointly have full positive control over the outflow.

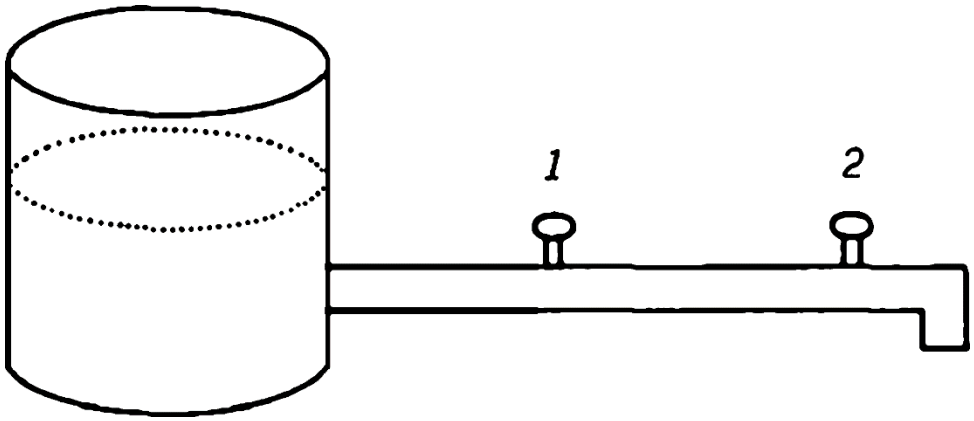

Figure 4.1: Reservoir and outlet pipe with shut-off valves at positions 1 and 2. Reproduced from Rescher (1969, p. 330).

With his distinction between full and partial control on the table, Rescher then distinguishes between full/partial control and influence:

> It is important to draw a distinction between partial control on the one hand, and what I propose—somewhat arbitrarily—to term *influence* upon the other. Essential in the idea of control is a condition of definiteness: the controller(s) can *definitely* make something happen or *definitely* preclude its happening. But there is also the prospect of what—by way of contrast with control—we shall call *influence*, viz., the capacity to make something's happening more likely or less likely. For example, the taking of vitamin pills may render it less probable that I shall contract a common cold. The pills do not give me control—not even incomplete control—over my catching colds: the connection is merely one of influence in the specified sense (i.e., taking the pill "influences" whether or not I shall catch cold). (1969, pp. 336–337)

Note that the capacity to make something's happening more or less likely corresponds to the idea of a dynamical constraint that was introduced in Chapter 2. In mechanistic contexts, *influence* corresponds to the presence of an active causal power possessed by that which does the influencing, and a passive causal power possessed by what is influenced.

Rescher uses various terminology to distinguish control from mere influence: words like *assure* and *definiteness*, the phrase "affect decisively" (1969, p. 345), and even the phrase *desired result*. These suggest that control is a kind of influence that rises above a certain threshold: a capacity for (or the exercise of) influence sufficient to warrant a certain level of confidence on the part of an agent who is interested in the outcome. Putting it this way

suggests that control may, like proper function, be relative to a normative *perspective*; but in particular, Rescher's terminology makes it sound like control is always relative to the (potential or actual) desires or goals of agents. This may make sense in the context of artificial control systems and when agents are themselves considered as controllers. But as noted in the introduction to this chapter, control is important in other contexts where it is not obvious that there is a particular agent that attributions of control should be indexed to. For example, control is exercised by parts of the nervous system of humans and other animals, at a level that is certainly below that of conscious goals or desires (Khoo, 2018). Increasingly, biologists explain cellular, subcellular, and even molecular phenomena in terms of control (Purich, 2010). One way to explain the normativity of control in such contexts is to say that it is nothing more than the normativity of proper function: i.e., that control merely refers to the kind and amount of influence that one thing has over another so long as it is functioning properly. In Section 4.5 I will argue that the normativity of control cannot be reduced to the normativity of proper function: many cases in which proper function implies a certain kind and amount of influence do not count as control, and many cases of control in biology are outside the scope of proper function. In short, the normativity of proper function and the normativity of control often come apart. A separate account of how the normativity of control is grounded is therefore needed to understand control in biology and how it is related to control in other contexts. In the next section I explain how control is implemented mechanistically as a type of observer-worker system that is distinct from the kind of observer-worker system needed to ground a standard of function/malfunction.

4.4 Autonomous Controllers as a Type of Observer-Worker System

Pattee argues that control only makes sense from a perspective on which there are multiple possible behaviors:

> in essence control implies that a system possesses *alternative* behaviors, and that owing to the particular nature of the constraint it is possible to correlate a controlling input variable or signal with a particular alternative output dynamics according to a rule. (1972, p. 83)

MacKay argued that we can only understand control, and the idea of there being multiple possibilities for a system, from a perspective that "under-specifies" the system, i.e., one that selectively discards a certain level of details:

> Only if we *under-specify* the situation, by using generic terms that allow more than one microstate to be compatible with our description, can we give meaning here to the word 'possible'. 'A 2-way switch is a device with two possible states...' Yes. But *this* '2-way switch' is a physical array of molecules in the *only* state compatible with its past history. 'Is no other state possible?'—Not in these terms. 'In future', of course, it is 'possible' that it may be found in either of the two states-but only if we have not specified the time, nor the intervening history of forces acting on it. Given these, the concept of a 'range of possibilities' evaporates. (1962, p. 93)

The fact that a controller must possess alternative behaviors (in other words, a discrete division of possible ways to respond to inputs) follows from the nature of controllers as "decisively" affecting an outcome or "assuring" a result. If one possibility is being assured, this means nothing more or less than that conflicting alternatives are assured not to happen. The point here is that control involves *selection* (Tolman, 1936, p. 253), and not merely influencing:

> A paradigm example of the contrast can be given in terms of a "fixed" roulette wheel in a gambling house. We should speak of "control" if the house can select the specific outcome of the wheel, but if the house can merely affect the probability-distribution of the outcome, without being able to determine any specific outcome, we should—in our terminology—have to speak of "influence" (rather than control). (Rescher, 1969, p. 337)

Selection involves an unequivocal and unambiguous *affirmation*, but also an unequivocal and unambiguous *negation* (Pitkin, 1912, p. 393). Since one discrete behavior is selected from a classification scheme, the other possibilities in that scheme are thereby unequivocally rejected (MacKay, 1969, p. 24). MacKay describes selection devices as working as if they were "key-operated" (ibid, p. 25), by which he means that a selection is meaningful only if there is a clearly defined range of possibilities the selection is being made *against.* In Section 4.2, I discussed the fact that all observer-worker systems have a discrete output (behavioral)

repertoire whose selections are executed in a context-dependent way,[55] and the fact that this is a necessary condition for anything that can ground a norm (recall Haugeland's statement that "norms have a kind of 'if-then' character, connecting sorts of circumstance to sorts of behavior," 1990, p. 151).

But do controllers fully count as autonomous normative systems? Do they count as observer-workers? For this to be the case, controllers need not only a discrete output repertoire but also a discrete input repertoire; they must be *observers* or *measurement devices*. In fact, the notion of selection itself implies that the system must be working with some sort of discrete input variable: there has to be a basis on which selections are made. But a system that simply generates random output behaviors is not, *by itself*, capable of making selections. A quarter can be flipped and will land one of two ways, and it can be used by an agent to make selections (in such cases the quarter would serve as the input to the selection process), but this does not mean that the quarter, by itself, can select anything.

Controllers also require inputs to meet the requirement of being able to *assure* an outcome. Assurance implies predictability, which implies regularity. The behavioral outputs of the controller at any given time must correlate, in some way, with some variable that could in principle be tracked separately from the outputs, from which the outputs could be predicted (and therefore, be seen as "assured" given the inputs). In order for a discrete output selection to be seen as "selected" with respect to an input, or as being "assured" with respect to the input, the input must itself be a discrete value with discrete alternatives. For these reasons, any system that is a controller will be an observer-worker system; controllers are a species under the larger genus of observer-worker systems.

[55] Rather than exhibiting Newtonian, or fully determinate dynamics, the linkage from inputs to outputs incorporates a selective loss of detail and therefore exhibits what Kelso refers to as a "pattern dynamics" (1995, p. 145).

But not all observer-worker systems are controllers. To see this it is important to consider another aspect that is analytically true of all instances of control: there must always be something doing the controlling, and something being controlled. As Rescher points out,

> Control over an object is never exercised over the object *as such* but over some *aspect* (feature, characteristic) of the object. Control is never unqualifiedly absolute. Control of something is always control over it *in respect to* such-and-such (its motions, movements, functioning, etc.). Nothing can ever be under one's control in *every* respect. Even when it is under his “full control” a driver cannot make a motor vehicle go sideways, upwards, etc. (1969, p. 348)

Rescher follows the general usage of engineers in referring to this aspect as the “control parameter.” It is, then, the presence of a control parameter that separates controllers from other types of observer-worker systems that are not controllers. Like observer-worker systems, controllers have a discrete output repertoire, but further, the potential outputs will be value assignments made to the same control parameter.

In the previous section, I remarked that the terminology Rescher uses to distinguish control from mere influence might seem to suggest that control is a kind of influence that is sufficient to warrant a certain level of confidence on the part of an agent who is interested in the outcome. I now claim that no such relativity to agent interests or mental states of agents such as confidence is essential to the existence of control. What is essential is that different conditions will result in discrete effects being produced by the controller on a control parameter. Such discrete operation may be such as to warrant a certain degree of confidence in an interested external observer, but this is incidental to the fact that control is taking place.

4.4.1 Autonomous versus Non-Autonomous (Mere Regulators)

It is important to clarify at this point that my use of the term ‘controller’ is not intended to map onto ordinary usage. Up to this point, I have been interested solely with *autonomous* controllers, that is, systems that count as a controller solely due to their own intrinsic

characteristics (independently of external perspectives). There are many other devices that, although not counting as autonomous controllers, can be *used as* controllers by agents, or that may be interpreted as such by agents. I will refer to these as non-autonomous controllers. In such cases, it is more correct to say that the *agent* is doing the controlling, and using the non-autonomous "controllers" for the purpose of control. From now on I will reserve the term 'controller' to refer to autonomous controllers, and I will use the term 'regulator' to refer to non-autonomous controllers.[56]

As an example, an analog thermostat that works by simple negative feedback may count as a regulator, but not a controller, in the present terminology. This may be true even if it functions *automatically*, i.e., without the supervision and intervention of an agent ("automatic" is importantly different from "autonomous" in this context). While it may be true that there is a regular causal relationship between the ambient temperature in the room where the thermostat is located and the effect that the thermostat has on the heating device for that room, if the thermostat does not have its own categorization scheme both for input temperatures and output settings applied to the heating device, the thermostat cannot be said to have its own fixed perspective (which would require selective loss of detail at a fixed level of abstraction; see Section 3.5.1) on what the temperature is or what the heating device should be set to. It will therefore not count as an observer-worker system, and therefore not as a controller. The person who designed, built, and/or deployed the thermostat for use as a controller is actually the one doing the controlling, not the thermostat itself.

Without such a distinction, there would be almost no limit on the range of systems that would count as controllers. By exerting a gravitational effect on the planets that results in regular orbits, the sun might be said to "control" the planets of the solar system. Instead, on my terminology, it is correct to say that the sun *regulates* the planets, but not that the sun *controls*

[56] Collier and Hooker (1999, fn. 2) distinguish between "control" and "regulation" in a roughly similar way.

them. Further, due to the normative aspects of control, as with proper function, an attempt to reduce control to regulation (a non-normative notion) will fall prey to generality objections similar to those discussed in Section 3.3. In the following section, I will elaborate on the normative nature of control and how it compares to the normative nature of function.

4.5 Degrees of Effectiveness, Degrees of Control, and Malfunction

There are several non-equivalent, orthogonal dimensions along which control systems are evaluated, but these are often not explicitly treated as separate. This does not usually cause problems when discussion is confined to specific philosophical or engineering problems pertaining to control (i.e., confined to trees rather than the forest). However, this fact is probably partly responsible for the lack of discussions that attempt to understand the forest of control as a whole.

For example, two evaluative dimensions of control that can come apart are the *effectiveness* of control and the *degree* of control. The degree of control has less to do with the controller itself, and more to do with how it is situated (think back to the example of the reservoir and outlet pipe with shut-off valves at two positions). Now consider a person who drives to a pub, has several alcoholic drinks, and then drives home. The degree of control that the person has over the car will not change (if it is granted that the car itself remains in the same condition, e.g., the steering system does not cease to function properly on the way home, and there is not a second person attempting to control the car). However, what is likely to change is the *effectiveness* of the control that the person has over the car. Effectiveness has to do with things like the fineness of discriminations that can be made in input measurements and output selections, as well as the speed and accuracy with which adjustments are made to the control parameter.

As another example to illustrate the difference between degree of control and degree of effectiveness, consider the following example from Rescher:

> Think of a wartime aircraft navigator who, as sole survivor, is frantically working the controls to get the hang of what they do so as to be able to bring the aircraft under control. There is no question that he “is in control of” the aircraft, since whatever the aircraft does is being done in response to his settings and resettings of the control apparatus. But until he masters the workings of the situation so as to be able to coordinate this control at his disposal with his purposes we would not say that he “has control over” the aircraft (1969, p. 335)

Degree of effectiveness can also come apart from the degree of malfunction. As discussed in Chapter 3, malfunction requires a point of view as to what counts as proper functioning and what does not. Depending on how a controller is being used, a certain degree of ineffectiveness may not count as malfunctioning at all. In other contexts, there may not be a standard of function/malfunction that is applicable to the controller.

Whether or not it even makes sense to talk about malfunction with respect to a controller depends on whether the controller is being given a “job” in a larger system. There would have to be something like an “alpha creature” (described in Section 3.5.3) that has a perspective on whether the controller is functioning properly in a given situation, and enacts a censuring response. It is conceivable that the controller could be its own “alpha creature”; the controller itself might have the ability to monitor its own activities and enact a kind of “censuring” response. Complex, adaptive controllers (e.g., that use reinforcement learning or Kalman filters) like this will be discussed in the next chapter (Section 5.6). What is important to see for now, however, is that an autonomous controller can exist even when there is no “alpha creature”. A system can ground a standard of control even when it does not ground a standard of malfunction.

If autonomous controllers, like all observer-worker systems, ground a type of normativity, then what kind of normativity are they capable of grounding? Only that type of normativity that directly is generated by the type of “if-then” rule (as Haugeland put it), fixed in its level of generality, that is tacitly put into practice by that system by itself. This would be

nothing more than a normative standard linking discrete input values with discrete control parameter values. In other words, certain control parameters are implicitly considered as "going with" or being "appropriate to" certain input values by the controller. This is the bare normativity of control; a minimal autonomous controller does not actually have its own perspective on degrees of control, degrees of effectiveness, or malfunction. Each of these requires a more complex perspective.

We can now revisit the definition I stated in the Introduction to this chapter:

> Control is influencing (or being disposed to influence) the right aspects of something (i.e., that which is "being controlled") in the right ways to the right extent by the right means at the right times, where what is 'right' in a given context may be determined either by an external perspective or by the system's own perspective.

By taking a perspective on which control parameter value should be selected under what conditions, the system realizes a normative standard on which aspects (the control parameter(s)) of the controlled system should be influenced to what extent by what means (this will be determined by the behavioral repertoire) at what times (this will depend on the input/output mapping). However, the controller does not thereby necessarily have a second-order perspective on whether this first-order perspective and its implementation yields control that is *effective*, whether it amounts to a lot of control or a little, and whether it is functioning "properly" or not. Each of these would require either additional recognition and response capacities that go beyond mere autonomous control, or an external observer.

Control is a complex topic because it is easy to lose sight of what kind of normativity is minimally associated with control versus what kinds of normativity are usually associated with controllers by external agents. But it is also a complex topic because there are many different kinds of controllers. For example, we have been discussing controllers so far that only have a single input and output parameter, but of course controllers can have many input variables and many output parameters. They can vary in degree of effectiveness, degree of control, and degree of malfunction along any of these dimensions. Controllers also vary according to the

complexity of how they are structured and of how they can change over time. Eventually our goal is to understand how control relates to agential concepts like goals, intentions, motivation, and cognition. This will require a look at a number of different types of controllers and how they are distinct.

4.6 Types of Controllers

4.6.1 Negative Feedback and Homeostats

Negative feedback devices (sometimes called “homeostatic mechanisms”) are one of the most common devices used in control systems. A negative feedback controller is a special case of control in which the measured quantity is compared to a reference value, also referred to as a “setpoint.” The difference between these is then used to determine what the control parameter output value should be. The word ‘negative’ in negative feedback refers to the fact that the control system responses attempt to negate the perturbations away from the reference value.

Two commonly cited examples are a thermostat used to control a furnace, and the Watt governor that is used to control the speed of a steam engine. The input to the thermostat is usually the ambient temperature of a room determined by a thermometer. This value is compared to a reference value (the temperature setting for the thermostat). If the ambient temperature is lower than the reference value by a certain amount, the furnace is turned on. Otherwise, it is turned off.

As described in Section 4.4.1, not all negative feedback controllers are autonomous controllers. Some are automatic but non-autonomous controllers (i.e., they either don’t actually take measurements, they don’t have a discrete repertoire of control parameter outputs, or both) that have continuous inputs or outputs and that approximate what an autonomous controller

would do. Even though such controllers are not “autonomous” in the sense defined here, they may yet be highly effective regulators. This is true of the Watt governor. In the case of the latter, the flow of steam from the engine turns the main shaft, which has swivel arms hinged to it that can swing up and down freely. The faster the shaft spins, the further upward the arms move due to centrifugal force. The arms are then dynamically linked to a valve on the steam outlet of the engine so that when the shaft spins faster, the steam flow is reduced to compensate for the increase in shaft speed and the shaft speed is brought more or less back to normal.[57] In other types of engines the arms may instead be dynamically linked to a fuel intake valve or throttle.

Since the operation is fully continuous, there are no points where the system selectively discards detail at a fixed level of generality, i.e., the system does not work with its own categorization scheme at the input (the point where shaft speed is converted to a “signal”, i.e., the swivel arm height) or at the output (the point where the changes in the arm height and connected linkage mechanism directly operate the valve setting). Further, there is no literal reference signal that the shaft speed or swivel arm height is actually compared to. An external observer (perhaps the designer or a technician whose job it is to calibrate the device) may decide that there is, in effect, an engine speed that the system is currently calibrated to maintain. But this will only be an idealization subject to varying interpretations. There is no objective basis for discriminating between when the system is operating within reference range versus when it is compensating for perturbations; these are categories that can only be imposed arbitrarily by an external agent. It is for this reason that the Watt governor is not an autonomous controller.

It should be noted that control does not necessarily involve negative feedback or homeostasis. Negative feedback is a specific type of mapping between input values and output

[57] See https://www.youtube.com/watch?v=ASII3HWTT4U for an illustration.

values, where the output values are usually intended (by the designer of the system) to have a compensatory effect on the input variable being measured. In other control systems, the control parameter may not be causally linked back to the input variable (this is sometimes called "open-loop control"). Consider a light switch that turns a light on or off. The light switch itself involves measurement; the physical switch position is measured and converted to an effectively binary signal. This signal then operates the on/off setting of the light itself (the control parameter). The switch itself is not responsive to whether the light is off or on; causation only flows in the other direction (of course a light sensor can be added to the system to turn it into a negative feedback controller, as is often used for street lighting or smartphone screen brightness levels).

4.6.2 Servomechanisms and Coordination

Up to now, the only constraints on the inputs and outputs of the (autonomous) controller we have been considering are that the inputs consist of measurements, and the various possible output values correspond to different values of a single control parameter. The distinguishing characteristic of controllers as compared to other observer-worker systems is the second requirement. A controller, then, maps input measurements onto a discrete, ordered set of values falling within a range along a continuously variable dimension—that is, onto discrete values of what Grush (2007) refers to as a *quasi-spatial manifold*.[58] By contrast, the range of

[58] Grush distinguishes between spatial and quasi-spatial manifolds as follows:

> … spatial information is straightforward; I can see that one point of light is between two others, that it is closer to one than the other; I can get similar information via touch. The primary contrast here is with quasi-spatial information. Many of the channels of information we receive through sensation are such that the ideas they occasion have features that can vary along one or more dimensions, but these dimensions are not genuinely spatial dimensions. Sounds can vary along the continuous dimension of pitch, and also along the continuous dimension of volume; colors can vary along three

outputs of other observer-worker systems may instead by values of what Grush refers to as a *punctate manifold*, a set of values "not naturally orderable along a dimension of variation" (2007, p. 416).

Coordination is a special type of control in which a) the input repertoire also consists of a quantization of a quasi-spatial manifold, and b) the controller implements a systematic, monotonic transformation from input values to output values.[59] In plain English, these are controllers where the output may be said to "track" or to be "guided by" the input.[60] Coordination is extremely important for any system that needs to control the movement of something so that it is responsive to the movement of something else. As we will see later in this dissertation, coordination is the basic organizational motif on which agency is built; it is the reason why agency is important for animals and robots.

Again, a light switch may be considered an extremely simple example of coordination, but a better example might be a digital volume knob. A sensor measures the rotational position of the knob and sends a digital signal to an amplifier driver board, which decodes the volume control signal and amplifies an audio signal to an extent proportional to the knob rotation amount. Again, non-autonomous versions of such a controller are also common, for example by means of a potentiometer (or rheostat), which transduces rotational movement into an analog control signal that is routed to the amplifier.

> continuous dimensions of saturation, hue, and brightness; a felt surface can feel more or less solid as it offers more or less resistance to pressure. I will call these qualitative continua "quasi-spatial manifolds." (2007, p. 416)

Here, I will simply consider quasi-spatial manifolds to be a larger class that encompasses what he calls spatial manifolds.

[59] This definition combines features of, but also diverges from, both that of Kugler, Kelso, and Turvey (1980) and that of Grush (2000). As Grush notes, coordination may apply to higher-order manifolds (called "order parameters" by Kelso, 1995) that result from the stabilization of features of other manifolds. A monotonic transformation between higher-order manifolds may correspond to a much more complex, non-monotonic transformation between degrees of freedom of a single behavioral unit (see Kelso, 1995 for examples). Coordination may involve the use of what are sometimes called "cognitive maps" (Golledge, 2010) or "orienting schemata" (Neisser, 1976, Chapter 6).

[60] Sometimes the words 'coupling', 'entrainment', or 'synchronization' are used instead.

A slightly more complex form of coordination that is common in robotics is that of a *servomechanism*. This type of controller combines two of the types we have discussed: coordination and negative feedback (Milsum, 1966). Instead of having a fixed reference value, as in the simplest forms of negative feedback, coordination is used to make changes to a variable reference value of a feedback controller. For example, one controller may coordinate the reference value of a second controller to the movements of a joystick, causing the second controller to seek different reference values depending on the position of the joystick. Milsum writes:

> the power steering of a ship operates mostly as a [fixed-setpoint controller] in transoceanic passage, but then as a servomechanism during rapid in-port maneuvering. (1966, p. 33)

4.6.3 The Many Ways Controllers Can Be Complex

Servomechanisms illustrate one of the many ways that sophistication can be introduced into a control system. Not only can controllers be combined into a larger controller (composite controllers, considered in the next subsection), but controllers can also have multiple inputs and outputs (I will call these *multivariate* controllers). Further, their input and output values do not have to be scalar but can also be multidimensional (I will call these *multidimensional* controllers).

4.6.3.1 Composite controllers: Serial control, shared parallel control, hierarchies, and heterarchies

Controllers have their effects on control parameters, where a control parameter is a quasi-spatial dimension of variation within the system being controlled. An example of shared parallel control was already discussed in Section 4.3. This occurs when two controllers are operating on two different control parameters of the same system simultaneously. Serial

control, by contrast, occurs when measurements are made of some aspect of the dynamics of one system that has been shaped by a controller, and these measurements are then input to a second controller. Closed-loop control (of which negative feedback is a special case) occurs when one or more controllers are arranged into a cyclical control series.

In some cases, multiple controllers arranged into a series are discussed in terms of "levels" of control, or even a "hierarchy" of control, where an earlier (or "upstream") controller may be viewed as "higher-level" (or perhaps even "lower-level", as with sensory information processing). Such metaphors will not be used in this discussion. Instead, two different kinds of hierarchy will be discussed: nested controllers, and hierarchies *of* control.

Nested control occurs when one controller contains another controller as a component. This is one way to implement a servomechanism: a feedback controller might be equipped with a component that uses an input sensor to calibrate the setpoint value. Nested control is extremely common in industrial engineering applications, especially digital electronic or software control systems in which it is feasible to have an indefinite number of nested controllers.

For our purposes, the more interesting types of complexity (partly because they are more widespread in biological systems) are hierarchies *of* control.[61] This occurs when one controller operates on a control parameter which is part of another controller, so that one controller directly modulates the functioning of another controller. There are a number of ways that a controller may be modulated:

- The output from one controller may itself be routed as input to another controller.

[61] Note that all of the uses of 'hierarchy' (and its cognates) in this section are narrower than the way Pattee (e.g., 1970) uses 'hierarchy' when discussing control. Pattee's term 'hierarchical control' often corresponds to 'autonomous control' in the terminology of this chapter (but not always; see Section 4.7.2). However, at one point, Pattee (1973a, p. 102) uses the phrase "*autonomous* hierarchical control" to express what he (later in that paper) refers to as *statistical closure* (and in later papers as *semantic closure* or *semiotic closure*), a concept not discussed in this chapter.

- The output from one controller may affect the shape of the function another controller implements from its inputs to its outputs.
- The output from one controller may change the nature of the input or output *repertoire* of a second controller.

In the first case, there are two important variations. First, the second controller may treat the output from the first controller as a continuously variable signal that needs to be measured. This is a true instance of hierarchical control. The second possibility is that the second controller uses the output from the first controller as an encoded signal. The first controller then might more appropriately be referred to as a *sensor*, and the second controller an *effector*. A third controller might also be inserted between the first two that accepts both an encoded input and an encoded output. It might appropriately be called an *information processor*. The encoded signals may then be viewed as *representations* with semantic contents, since they are part of a larger autonomous control system that gives such encoded contents meaning.[62]

The second and third cases are treated in the following two subsections.

4.6.3.2 Variable controllers: Non-static output repertoires and input/output mappings

Above, I mentioned that controllers can also have multiple inputs and outputs (these are *multivariate* controllers), and that input and output values of controllers do not have to be scalar but can also be multidimensional (*multidimensional* controllers). These should not be confused with a much more radical form of complexity, that exhibited by what I will call *variable* controllers. A variable controller is a controller in which either the input/output mapping or the output repertoire (or both) can change over time.

[62] A general account of the nature of representation, semantic information, and content is beyond the scope of this dissertation. For such an account that seems to fit well with the present account of control, see MacKay (1969). In my view, MacKay's account of how to naturalize contents remains unsurpassed and vastly underappreciated—a "wheel" that many others have failed to successfully reinvent. For an argument that the conditions I have stated in this paragraph are necessary and sufficient for representation, see Pattee (1970).

One simple example of the output repertoire being variable in a controller is a music synthesizer keyboard with a pitch bender wheel. The pitch bender is a wheel on some synthesizers that can be rotated to different positions; changing the position of the wheel changes the pitch of all the keys of the keyboard. The input/output mapping remains the same, but the behavior produced by output repertoire selections is changed.

A simple example of the input/output mapping being variable in a controller is a computer keyboard with a switch allowing it to be set to QWERTY or Dvorak mode. The output repertoire remains the same (either mode leads to the same set of possible character sequences), but the input/output mapping is changed; elements of the input repertoire become mapped to different elements of the output repertoire.

Variable controllers often have certain inputs that can act as a “mode switch”, as with the computer keyboard. In animal species, this can take the form of instincts and emotions. An animal may recognize a certain stimulus pattern, corresponding to the presence of a certain kind of situation, that then stimulates readiness for certain types of action, or leads to a change in the animal’s dispositions for certain kinds of behavioral responses (Frijda, 2007).

4.6.3.3 Goal-Directed systems

McFarland (1989) usefully distinguishes between goal-achieving systems, goal-seeking systems, and goal-directed systems. A goal-*achieving* system is “one which can recognize the goal once it is arrived at (or at least change its behaviour when it reaches the goal), but the process of arriving at the goal is largely determined by the environmental circumstances” (1989, p. 108). This would seem to include any controller that can change its behavior (for example, change its output repertoire) once a certain condition is detected. A goal-*seeking* system is merely any system that progresses towards a goal as a result of its own organization or design. Neither of these notions is particularly useful for characterizing types of autonomous

controllers, since they are both characterizations that may potentially depend on the perspective of an external observer.[63]

Goal-*directedness* will instead be the focus of this subsection. I will divide goal-directedness into weak goal-directedness and strong goal-directedness.[64] Weak goal-directedness involves *targeting*, but not necessarily model-based control. Targeting is a special type of coordination, in which two controllers are connected such that the first modulates the second, such that the modulation counts as coordination. Such modulation is then called *aiming*. When aiming is performed continuously overtime, and combined with negative feedback, the resulting type of control may be called *guided targeting* (Klinger, 1977, p. 84). Usually guided targeting has the effect of reliably causing the controller to ultimately have some effect on a specific object or location in the environment, and the presence of this object is what produces the measurements that the coordinator is "tracking". In this case, the object or location may be referred to as the "target", and the targeting system is "locked onto" it. The object may also be referred to as the "goal" of the controller, depending on what type of effect the controller will ultimately have on the object.

I refer to this as "weak" goal-directedness because although guided targeting and tracking are objective phenomena when performed by an autonomous controller (for example, a guided missile), there is often no literal goal representation in the system (except in the sense that measurements are being taken that effectively track the location of the goal). The ultimate action that will be taken with respect to the goal (if any) is also not explicitly represented within the system itself; nor are the consequences of such actions nor any kind of explicit valuation of

[63] Unfortunately, many authors in many different fields often use the term "goal-directed" to mean either what McFarland means by "goal-seeking" or "goal-achieving," or teleological in some other sense (for example, Murphy and Brown, 2007, p. 106 write that "all activity [of organisms], even of the most rudimentary sort, is *goal-directed*"). Below, when I use the term "weak goal-directedness", I will still mean a stronger sense of "goal-directedness" than these notions. "Weak" and "strong" goal-directedness is defined in the next paragraph.

[64] McFarland's understanding of goal-directedness corresponds to what I call *strong* goal-directedness, defined below.

such consequences. All of these features would require model-based control (to be explained shortly). However, I do not consider all of these features to be necessary for strong goal-directedness.

Although weak goal-directedness is not sufficient for getting a grip on model-based control or the use of explicitly represented goals, it is a very important form of goal-directedness because it is the simplest way that a system can have a perspective on which its behavior is directed at a robust object (not merely a control parameter). Weak goal-directedness will therefore serve as a crucial foundation for understanding desire-based motivation in the next chapter (Section 5.3.3).

4.6.3.4 Model-based controllers and goal-directedness

There are many ways that models can be incorporated into control systems; here, I will only be able to focus on a few of these. First, there is a sense in which any autonomous controller may be said to involve a model. Pattee writes:

> There are many relatively simple biological recognition-action structures that might suggest a primitive kind of model. For example, seedlings detect gravity and light, and by converting these input observables to specific rates of growth they control their morphology. A physiologist might prefer to call such tropisms a stimulus-response action and reserve the concept of model for a more complex relation between recognition and action. A cybernetician, on the other hand, would claim that the seedling has a model of its world, however primitive (Ashby 1956). In higher organisms we can recognize the nervous system as the physiological structure with the primary function of mapping sensory inputs from various receptor organs to output actions of muscles, and we often restrict the idea of model to mappings or representations in the brain. However, in the context of the more or less gradual process of evolution we do not learn much about primitive necessities for function by looking only at highly evolved organisms. There is generally more explanatory power in studying the origin of functions. What are the minimum requirements for this modeling relation in organisms?
>
> An engineering description of a modeling relation would include at least three functions:
>
> 1. Detection, recognition, or measurement that transforms a physical pattern into model inputs. In organisms these are usually called receptors or sensory organs;
> 2. the model itself that establishes the particular input/output relation; and

3. the effectors that are controlled by the output of the model, and that interact again with the physical environment. (1996, p. 255)

Pattee's application of the notion of a *model* shown in this quotation explains why he often claims that control systems in biology "contain their own descriptions" (e.g., 1971, p. 265). The constraints that enable the autonomous controller to act the way it does essentially conform it to an implicit model of its own behavior: "The enzyme molecule is a set of co-ordinated constraints which classifies its collisions with other molecules—the only sensitive collisions being with the substrate" (Pattee, 1971, p. 272).

Similarly, Pepper argued that the simple "chain reflex" control system of the digger wasp (wasps of the genus *Sphex*) constituted a kind of model:

> The world is categorized for the wasp as a sequence of causally connected events from the appearance of a grasshopper to a well-stored hole in the ground. The wasp does not wonder at the miracle of how nature should conform to these special categories. … The wasp just acts and on the whole perpetuates her species. The wasp's chain reflex categories, of course, were generated by an environment which had the regularities to support them. (1958, pp. 107–108)

Pepper argues further that in general, the motivational systems of animals may usefully be looked at as involving models of the environment that he refers to as "anticipatory sets". Motivational systems, like hunger or thirst, are activated by what Pepper refers to as "impulse patterns"; for example, thirst may be activated by the organism's detection of a condition of dehydration. The organism will become more sensitive to cues in the environment related to opportunities to drink water, and will experience a heightened readiness to respond to such cues in certain ways. For simpler organisms like the digger wasp, a sequence of highly specific anticipatory sets will be triggered, each prescribing specific responses to environmental cues—as if from a script—that triggers the next anticipatory set in the sequence, producing what Dennett (1984, p. 11) refers to as "sphexish" behavior, or what Sterelny (2003, p. 18) refers to as "detection cascades."[65]

[65] Keijzer (2013) argues that the digger wasp's behavioral capacities turn out not to be quite as "sphexish" as Dennett makes them out to be.

More sophisticated animals are not limited in these ways, but psychologists and neuroscientists generally rely on the notion of *motivational systems* to explain behavior (Toates, 1986; McFarland & Kalivas, 2003). Herbert Simon (1979, pp. 4–5) argued that motivational systems are necessary for the cognitive system of any "creature of bounded rationality"—including humans—to impose constraints on relevant sensory information, possibilities for action, and problem solving strategies that will be tailored to the specific features of the situation. According to Kenrick et al.,

> any motivational system includes (a) a template for recognizing a particular class of relevant environmental threats or opportunities, (b) inner motivational/physiological states designed to mobilize relevant resources, (c) cognitive decision rules designed to analyze trade-offs inherent in various prepotent responses, and (d) a set of responses designed to respond to threats or opportunities represented by the environmental inputs (i.e., to achieve adaptive goals). (2010, p. 306)

Similarly, Neisser defined motives in terms of schemata:

> The activities of schemata are not contingent on any external sources of energy. If the right sort of information is available, the schema will accept it and may direct movements to search for more. But organisms have many schemata, related to each other in complex ways. Extensive schemata typically have less wide ones embedded in them…. When they do, the larger schema often determines, or "motivates," the activity of those embedded within it. Motives are not alien forces that bring otherwise passive systems to life; they are just more general schemata, that accept information and direct action on a larger scale. (1976, p. 56)

Motivational systems are generally activated by the recognition of a certain type of situation (for example, thirst may be activated by the sensation of dehydration); they then sensitize the organism to certain affordances in the environment. The thirsty animal will see the environment as parceled into things based on their relevance to opportunities for drinking, and its action selection system will be primed for relevant behaviors. Once that need is satiated, the animal will become less sensitized to drinking opportunities, and other motivational systems may become more dominant based on their triggering conditions. Humans have motivational systems not only for basic physiological needs like hunger and thirst but also higher-level "needs" such as achievement and affiliation (Schultheiss & Brunstein, 2010, p. xix).

Motivational systems are often discussed as being systems that “control” certain kinds of behaviors. On this way of speaking, motivational systems could themselves be considered as separate control systems, with their own input repertoires (set of affordances), input/output mappings (behavioral response strategies), and action repertoires (Frijda, 2007). But we can also look at them as being part of a larger, complex *variable* motivational system that includes in its input repertoire the set of triggering situations that will activate particular motivational subsystems. In this way, the entire motivational system, including its subsystems, constitutes a model that any given organism will use in its species-typical interactions with the environment.

There is another important sense in which control systems can involve models. So far, nothing has yet been said about learning and behavioral plasticity, which is a crucially important aspect of animal behavior. In fact, in control theory, as well as in discussions of animal behavior, the phrase “model-based control” generally refers not to implicit assumptions that are hard-wired into an animal’s nervous system or an artificial controller’s circuitry, but instead to the use of models of the environment that can be updated based on experience.

It has often been pointed out that these latter types of models, which are often referred to as “representational,” are not necessary for certain kinds of learning and adaptive behavior (Arkin, 1998; Krichmar, 2012). Pavlovian conditioned response is a type of learning mechanism in which input/output mappings between sensory inputs and behavioral responses are modified based on detected correlations between input values. Highly sophisticated non-representational adaptive control systems have been developed for robots that use neural network-based learning algorithms. In some cases, the latter make use of something like *reward* or *reinforcement*, in which a utility value is assigned to certain detected events. Arguably, a system that makes use of a reward or utility function to modify its behavior is now in the territory of what we might call “strong” goal-directedness. The presence of reward, utility, or error-detection indicates the presence of a standard of success and failure that is implicitly in use by a particular motivational system for the purpose of generating behaviors.

An even “stronger” form of goal-directedness combines what I have called “weak” and “strong” goal-directedness into a system that is capable of tracking and seeking an object or condition in the environment and using trial-and-error strategies to acquire or consume the object, and to remember those strategies that “work” and those that do not. Again, systems like this do not necessarily require the use of an anticipatory model for the purpose of generating strategies (Dayan, 2012; Coutlee & Huettel, 2014), but some have argued that animals such as rats do use such models and select strategies based on an evaluation of the predicted outcome (Dickinson & Balleine, 2002; O’Doherty, Cockburn, & Pauli, 2017).

To summarize, I have considered three main ways that control systems may be said to involve models (again, this is not an exhaustive list):

- Any autonomous control system that maps sensory inputs to behavioral outputs constitutes an implicit model prescribing how the overall system will behave (the Pattee sense).
- The use of a model to maintain information about the environment, including features of the environment that are not currently being detected or perceived (*forward models* are commonly used in this way).
- The use of a model to predict the outcome of hypothetical events, behaviors, or strategies.

I have also at this point considered several forms that goal-directedness can take:

- Tracking a detected object in the environment, by means of coordination between multiple control systems (one that detects changes in the object’s position, and one that is coordinated with the first that repositions the sensory or motor system to “follow” the object). This is “weak” goal-directedness.

- Internal assignment of a utility, reward, or error value to detected outcomes of behaviors that is used to make changes to the input/output mapping. This is "strong" goal-directedness.[66]
- Combination of weak and strong goal-directedness to guide and refine strategies for acquiring or consuming (or avoiding or destroying, etc.) an object or type of object. This is "stronger" goal-directedness.

There is a form of goal-directedness that is even stronger than the ones listed above, in which a standard of utility/reward is adopted as a form of "common currency" across multiple motivational systems to resolve behavioral conflicts (McFarland & Sibley, 1975). This will be discussed in the next chapter.

4.6.4 Metamorphic Controllers: Non-Static Input Repertoires

A controller is metamorphic when its input repertoire can change over time. A simple example of this is a digital camera that can be set to different resolutions. I refer to these as metamorphic because it is really with the ability to change input repertoires that a controller gains the ability to radically change in nature. Neural networks are sometimes used to implement unsupervised learning algorithms for radically metamorphic controllers that can discover new kinds of patterns or affordances that they were not previously sensitive to; they thus increase the variety of patterns the system can respond to (i.e., its input repertoire). Some metamorphic controllers can dynamically alter the processing resources, input resolution, and localization of input sensitivity, simulating or recreating biological perceptual phenomena such as focus, attention, and concentration.

[66] Note that this goes beyond mere negative feedback control, since feedback is here being used to make changes to the input/output mapping itself.

Motivational systems and their internal states can alter the kinds of input patterns an organism or agent is sensitive to, as well as the way attentional resources and concentration are allocated. Automatization and the development of motor skills and sensorimotor coordination can also serve to alter one's input repertoire to become much simpler and optimized for task performance. The human mind is also a metamorphic controller in that it can undergo different *mindsets*. According to Reeve,

> A mindset is a cognitive framework to guide one's attention, information processing, decision making, and thinking about the meaning of effort, success, failure, and one's own personal qualities. Once adopted, a mindset functions as a cognitive motivational system that produces many important downstream consequences in one's thinking, feeling, and acting. That is, the person with one mindset looks at a motivational episode in a fundamentally different way than does the person with a different mindset, and these different ways of thinking yield differences in lifestyle and ways of coping. (2015, p. 240)

Mindsets can be consciously adopted or triggered automatically, which especially happens when one begins striving for a particular goal (Braver et al., 2014). Conceptualization and cognitive priming are additional important factors that affect the input repertoire of cognitive and motivational systems.

4.7 Summary and Comparison to Other Accounts of Control

In this chapter I have offered an account of control on which a controller puts into practice a perspective on what kind of influence (what changes should be made to which control parameters) is appropriate under what conditions. Control is influencing (or being disposed to influence) the right aspects of something (i.e., that which is "being controlled") in the right ways to the right extent by the right means at the right times, where what is 'right' in a given context may be determined either by an external perspective or by the system's own perspective (here using 'perspective' in the sense explained in the previous chapter). Control perspectives can be extremely simple (as in the case of a light switch) but can also be as

complex as the human mind itself. I have offered a taxonomy of control systems to provide a basic framework for understanding how notions like goal-directedness, coordination, attention, and motivation relate back to control.

The most fundamental distinction to make about controllers is that between autonomous and non-autonomous controllers. An autonomous controller works with its own categorization scheme for input and control parameters and maps input categories to control parameter values. A non-autonomous controller (or “mere regulator”) does not have its own perspective in this sense, but may be counted as a controller by an external observer who defines its input and control parameters. Variable and metamorphic autonomous controllers also work with a perspective, but it is a perspective on which the standard for what kind of influence is appropriate under what conditions can change in highly complex ways.

In the following subsections I discuss how the present account compares to other attempts at understanding the nature of control.

4.7.1 Rescher

As discussed in Section 4.3, Rescher’s position is that “Positive [full] control involves the power to assure a desired result, negative [full] control the power to prevent an undesired result” (1969, p. 331). We can interpret his use of the word ‘assure’ within the framework adopted in this chapter. When an input parameter (or parameters) of an autonomous controller takes on a certain setting, an output parameter setting (or settings) will be *selected* as a result. The fact that autonomous control works by *selection*, rather than mere influence, means that *within the controller’s own implicit framework*, one output parameter configuration will be selected and all others rejected. The controller will have its own implicit categorization framework, or *perspective*, if it discriminates differences in parameter values at a fixed level of generality, which will involve a selective loss of detail (Section 3.5.1). The fact that autonomous

controllers can “assure” a certain result in given cases is exactly what makes them useful for so many applications.

I have already discussed reasons why Rescher’s invocation of the notion of ‘desire’ is too anthropomorphic; many biological control systems work at a scale where it would not make sense to talk about desires, and they exist in organisms that are arguably too simple to have desires (at least, not in a sense that is uncontroversially non-metaphoric). But why does Rescher appeal to the notion of ‘desire’ in his account of control?[67] Though it may not be true that a simple autonomous controller (e.g., a light switch) literally desires a certain output parameter setting, it may make sense to say that the controller grounds a standard of what response is *appropriate* given a certain input. Of course, this is an extremely limited standard of appropriateness; its scope applies only to the controller itself, and my have no relevance or significance whatsoever beyond it. For a light switch, i.e., *from its* (almost trivially simple) *perspective*, when the switch is in the on position, that is the time when it is appropriate for the light to come on. This is a standard of control, but not a standard of function/malfunction. If the light switch ceases to function properly, it ceases to ground the standard on which the light’s coming on is appropriate to the switch being in the on position. Only from the perspective of an external observer would that standard still be applicable to the system in any way (such as in the judgment that “That is a malfunctioning controller”).

The invocation of a notion like ‘desire’ may have plausibility because we tend only to identify normative control standards in contexts where they may fulfill the desire of some external agent—where they may potentially provide opportunities for purposeful *agential* control; in other cases, the operative standard simply isn’t salient. At one point, MacKay argues that this type of consideration underlies the very concept of control itself: “the subtle and arbitrary human element that underlies many cybernetic notions [is that] basically, by saying

[67] Similarly, Dennett argues that “for something to be a *controller* its states must include desires—or something ‘like’ desires—about the states of something (else)” (1984, p. 52).

that A controls B we mean that if *we* could control A then *we* could control B" (1964, p. 311). While potential desirability correlates well with control standards that interest humans, and may even enter into the way people think about or conceive the notion of control, this does not mean that potential desirability is a necessary condition for the existence of control.

4.7.2 Pattee

Pattee attempted to define a notion of control that is highly general, applying it even to individual enzyme molecules. In general, Pattee was only interested in what I have called *autonomous* control, for example writing that "control can only arise through some selective loss of detail" (1973a, p. 95). He often writes about "hierarchical control" but it's not clear that he intended this as a particular class of control, to be contrasted with "non-hierarchical" types of control. It appears more likely that he took there to be a control "hierarchy" involved in any case of control. However, there seem to be two distinct senses of "hierarchical" when Pattee discusses control (though he does not seem to have made this distinction explicit). Anytime Pattee writes about control, he discusses it in terms of a hierarchy of levels of description, due to the fact that control involves a selective loss of detail, in which the dynamics of a system are constrained so that only a limited set of possibilities are open to it. The controller makes a selection from among this limited set of possibilities based on the input parameter:

> it is important to realize that *controls must operate between different descriptive levels,* just as constraints must be defined by different descriptive levels. This is necessarily the case for all measurement, recording, classification, decision-making, and informational processes in which a number of alternatives on one level of description is reduced by some evaluative procedure at a higher level of description. Why are these necessarily two-level processes? Why are two distinct descriptions necessary? Because we cannot speak of an event as being both possible and impossible using the same level of description. On the lower, unconstrained level the alternatives must be possible; for if they were impossible then deciding for or against them would be a vacuous process. But on the upper, constrained or controlled level, in so far as the rules are reliable or effective, some of these alternatives are actually selected (Pattee, 1972, p. 84)

On the other hand, Pattee sometimes describes “control hierarchies” as systems that not only have a constrained set of behaviors, but also in which both hierarchical levels act to constrain each other:

> In a control hierarchy the upper level exerts a specific, dynamic constraint on the details of the motion at lower level, so that the fast dynamics of the lower level cannot simply be averaged out. The collection of subunits that forms the upper level in a structural hierarchy now also acts as a constraint on the motions of selected individual subunits. This amounts to a feedback path between levels. (1973a, p. 93)

In cases such as Pattee is describing here, the “control hierarchy” involves a cycle of control: one process constrains the behavior of the system by means of selective loss of detail, as described in the 1972 quotation. But the resulting constrained behavior then leads to effects that *are* sensitive to the finer level of detail selectively ignored at the earlier step. Pattee argues that such inter-level processes (occurring *between* descriptive levels) that feedback onto each other are crucial for the functioning of biological systems, and are also the most mysterious aspect of them. As noted above, Pattee often refers to this latter situation as “statistical closure”, “semantic closure”, or “semiotic closure”. Pattee's stronger notion of a "control hierarchy" is concerned with the origin and distinctiveness of living systems, a topic beyond the scope of this dissertation. Pattee’s weaker notion of a “control hierarchy” (described in the 1972 quotation) is more relevant here: it is slightly more general than my notion of an autonomous controller, and is equivalent to what I have called an observer-worker system.

Again, the difference between an observer-worker system and an autonomous controller is that autonomous controllers are a *species* of the wider genus of observer-worker systems. I argued above that not all observer-worker systems are controllers. Recall Rescher’s point that “Control of something is always control over it *in respect to* such-and-such (its motions, movements, functioning, etc.)” (1969, p. 348). It is the presence of a *control parameter*, that is, some common dimension of variability that the output behaviors influence in different ways, that separates controllers from other types of observer-worker systems that are

not controllers. Like observer-worker systems, controllers have a discrete output repertoire, but further, the potential outputs will be value assignments made to the same control parameter.

4.7.3 MacKay

MacKay offered a criterion of control that is endorsed by various authors from time to time, arguing that

> "Control" not only implies corrective reaction, but the conceptual possibility of its *absence owing to lack of information*. Unless there is a separate information-path which could conceivably be interrupted, the concept of control is inapplicable and the reaction could perhaps best be described as "Newtonian". (1952, pp. 55–56; cf. Kelso, 1995, pp. 144–145)

In a later article, MacKay elaborated on what he means by an "information-path":

> The only objective physical distinction we can firmly draw is between (a) devices, such as watercocks, steam-valves, transistors and rudders, where the input, A, determines the form of the output, B, without supplying all the energy of B; and (b) devices such as transmission lines, levers, springs and gear trains, where the energy of B is totally provided from the energy of A. In the first case, the energy of A is at least partly devoted to altering the structure through which the energy for B is channeled—altering the coupling between the output, B, and its internal energy supply. In the second, no analogous process occurs. In the first case a cybernetician would say that A exerts 'active control' over B. In the second (if we wish) we may speak of 'passive control'; though to some of us it would here seem clearer to speak simply of action and reaction. The important point is that in cybernetics we are concerned with the action of *form* upon *form* rather than of *force* upon *force*. (1964, pp. 311–312)

Because I have argued that control requires an observer-worker system, this implies that there will always be an instance of *recognition* (in the Dretske sense; Section 3.5.1) at the point of input and behavioral *selection* at the point of output. For Dretske (and Pattee, and others) what is important for recognition is *selective loss of details*. Dretske argued that selective loss of details makes it possible for a system to be sensitive to whether or not multiple tokens count as instances of the same general type. Another way of putting this might be that selective loss of details makes a system sensitive to whether a certain *pattern* or *form* is present.

However, this does not mean that the full energy for the control process cannot be supplied by the input. Consider, for example, a bicycle gear shifter with discrete input settings. To switch to the next gear, a certain amount of force must be applied that reaches a certain threshold. Before the threshold is met, the derailleur does not change position. Once enough force has been applied, the derailleur switches position, moving the chain to the next gear. But it is the energy applied to the shifter that actually *moves* the derailleur, and that therefore moves the chain to the next gear. On the account of control offered here, this example would count as an autonomous controller even though the controller's input supplies all of the energy for the output. Contrary to MacKay, even in autonomous controllers there need not be a separation between the energy path and the information path.

4.7.4 Shepherd

Shepherd (2014) claims to offer a "general account" of the "nature of control itself." While Shepherd's goal is to offer a broad account of control, his main motivation is to offer an account of control that explains what is lacking in cases of deviant causal chains, the most common type of counterexample that has been raised against Davidson's causal theory of action.[68] Philosophers of action who wish to defend the causal theory of action and solve the problem of deviant causal chains are generally not only interested in the question of when does control *exist*, but also the question of under what conditions control is *effective,* under what conditions a sufficient *degree* of control exists to say that the causal chain is not deviant, and under what conditions an instance of control has or has not *malfunctioned*. When one looks at the range of thought experiments that are treated in the debate, it becomes clear that an insufficient degree of control, an insufficient degree of effectiveness, and malfunction are all

[68] The causal theory of action and problem of deviant causal chains are discussed in Davidson (1973).

ways that a causal chain can be "deviant." The deviant causal chain debate is complex mostly because these three dimensions of normativity that can be associated with control—explained above in Section 4.5—are often conflated in that debate.

In this dissertation, I am not offering a solution to the problem of deviant causal chains, and offering an account of how to determine or quantify the effectiveness of control or the degree of control in particular cases, or what constitutes malfunction with respect to agential control, would all be outside the scope of this dissertation. Instead, I want to look at Shepherd's account merely in terms of how it characterizes the basic nature of control and how it compares to the present account in that respect. Shepherd's account is as follows:

> An agent J exercises control in service of an intention I to degree D in some token circumstance T if and only if (a) J's behavior in T approximates the representational content of I to (at least) degree D, (b) J's behavior in T is within a normal range for J, where the normal range is defined by J's behavior across a sufficiently large and well-selected set of counterfactual circumstances C of which T is a member, (c) the causal pathway producing J's behavior in T is among those normally responsible for producing J's successes at reaching the level of content-approximation represented by D across C. (2014, p. 410)

We can set aside condition (a) since it is specific to *intentional* control. Intentions will not be discussed until the next chapter, but essentially an intention is a specific type of goal (in the sense of strong goal-directedness). I have argued above that the input/output mappings of controllers need not involve goal-directedness in either the weak or the strong sense. Condition (b) is somewhat similar to the condition on my account that control involves the selection of a behavior from the controller's behavioral repertoire.[69] Condition (c) is the operative condition that is designed to meet causal deviance challenges. Translated into the more general concepts of the present account, it might read as follows: "the causal pathway producing the controller's behavior is among those normally responsible for producing the output behavior

[69] The appeal to counterfactuals is of course a departure from the present account; see Section 3.3 to see why I have rejected counterfactual accounts of the normativity of proper function. Similar considerations apply to the normativity of control.

corresponding to the input according to the input/output mapping of the controller." Shepherd clarifies that "the notion of normal here is statistical" (2014, p. 407).

The main weakness that I see with condition (c) is its reliance on statistical normalcy. The account implies problematically that the first time that a controller that has just been built is put to use cannot count as control because since there have been no previous cases of control, no causal pathways are yet "normally responsible" for producing the behavior. Instead of understanding the causal pathways within a controller in terms of statistical normalcy, I propose instead that we understand them in terms of the notion of a causal pathway explicated by Ross (forthcoming-b). According to Ross, scientists such as biologists often treat causal pathways not as statistically normal processes, and not as mechanisms, but instead by means of a concept that

> captures a (i) sequence of steps, where these steps (ii) track the flow of some entity [such as matter or energy] through a system, (iii) abstract from significant causal detail, and (iv) emphasize the "connection" aspect of causal relationships. (forthcoming-b, p. 6)

Ross points out that pathways often do not reduce to mechanisms, because they can be realized by multiple mechanisms, i.e., matter and energy may flow through the same pathway by means of different kinds of mechanisms at different times. This is especially common in sophisticated control systems such as in the brain. However, as with mechanisms, any given causal process running through a pathway will consist of an unbroken chain of causal influence from one component to another; ultimately this causal influence will be conducted by means of directional causal powers that emerge from second-order constraints (Section 2.4.5).[70] Input/output mappings in controllers should therefore be seen as implemented and individuated by causal pathways in Ross's sense. For purposes of addressing deviant causal chains, Ross's pathway concept can provide the flexibility needed to account for multiple means of causal

[70] Winning and Bechtel (2018) discuss the fact that the causal powers involved in control will often be implemented by means of non-holonomic constraints. The terminology of holonomic versus non-holonomic constraints was defined in Section 2.3.4.

influence within the same controller but is an account of causal pathways that is grounded in the present, actual nature of the control system, rather than counterfactuals or its statistical track record.

4.7.5 Ross

In another paper, Ross provides an account of the nature of control that is intended to account for *causal selection*, the "distinction we make between background conditions and 'the' true cause or causes of some outcome of interest" (forthcoming-a, p. 1). She argues that causal selection is important "in the context of biomedicine, where scientists commonly identify 'the' cause or causes of specific diseases" (ibid.), and that in those contexts, such identifications are made on the basis of the "*causal control*" that candidate causes exert over the disease in question. Much of her paper is then devoted to explaining the nature of causal control as it applies in such contexts.

Two important reasons why determining the causes of diseases involves identifying controlling factors are that "diseases are (1) type level phenomena, which (2) are often represented as taking on the values of 'present' or 'absent'" (ibid, p. 4). First, by "type level phenomena", Ross means that we often count multiple cases that differ to some extent in their details to be instances of the *same* disease type. The attribution of sameness is on the basis of specific causal etiology, i.e., there is a set of common "cause and effect *variables* that participate in a type-level causal relationship" that "produce all or most instances of disease D" (ibid, p. 8). Second, Ross argues that in given cases, diseases are generally considered to be either present or absent; the presence of a disease is treated as a discrete, binary variable.

Ross argues that these two aspects of diseases explain why *control* is relevant for the discovery of causes. The type-level and discrete, binary aspects of diseases lead to a *contrastive*, counterfactual understanding of disease causes:

> **(i.c.) interventionist cause:** a factor C has causal control over disease D if and only if there are circumstances S such that if some (single) intervention that changes the value of C (and no other variable) were to occur in S, then the value of D or the probability distribution of D would change, *for the contrastive focus in question*. (ibid, p. 5)

Ross notes that i.c. "involves a counterfactual claim: it maintains that C has causal control over D in the sense that *if* there was a change in C, this *would* produce a change in D" (ibid.).

We can make sense of why i.c. involves *control* by recalling Rescher's distinction between influencing and selection: C does not merely influence D but instead *selects* D. In this sense, an organism's responsiveness to the factors that control whether or not it acquires the disease amount to an observer-worker system, an autonomous controller, with its own discrete input repertoire (presence or absence of the intervention) and output repertoire (presence or absence of the disease). Ross's account of control is therefore largely compatible with the one developed here.

However, to truly be a case of control, more than just a counterfactual dependence should exist between input and control parameters: I argued in the previous subsection that the dynamical linkage between input and output should be considered as implemented by means of a *pathway* of dynamical *constraint*. In Section 2.3.5, I argued that constraints are not mere occurrent regularities, but that they are the *modal facts* about a dynamical system. Constraints in a dynamical system pertain not only to what does happen but to what *might* happen; they are the truthmakers for dynamical equations and modal causal claims. The account developed here, of autonomous controllers that are constituted by constraints capable of a selective loss of details, can therefore support a non-external-perspectivalist account of causal relations that admit of discovery by means of contrastive causal selection.[71]

[71] In another paper under development, I argue that in general, causal connections that operate within the scope of nomological machines are amenable to interventionist types of causal explanation and causal control precisely because of the fact that a nomological machine *just is* a type of observer-worker system.

4.8 Conclusion

The goal of this chapter has been to offer a framework for understanding control that sheds light on what it is about control that makes it useful in so many contexts. This consists in the fact that control involves selection rather than influence, and controllers can autonomously (i.e., independently of external perspectives) realize a normative standard of which control parameter value selections should be made in which situations. While all autonomous controllers share these traits in common, they can differ massively from one another in terms of complexity.

Part of what makes it difficult to understand how discussions about control in disparate contexts and subjects relate to one another is that multiple normative dimensions apply to control. Autonomous controllers have a normative perspective on which parameter values to select in which situations. But external observers also evaluate controllers in terms of the *degree of control* that a given controller may offer over some other variable, in terms of the *effectiveness* of a controller with respect to some control variable, or the *degree of malfunction* that might apply to a controller. It is important to keep clear that autonomous controllers do not, in general, have a built-in implicit normative standard that applies to these other dimensions, but this does not diminish their autonomy, i.e., their nature as being a controller (and not just a mere regulator).

The key to understanding how concepts usually associated with agency such as goal-directedness, motivation, attention, etc. are related to simpler control systems lies in understanding the different ways that control systems can be complex. As I will argue in the next chapter, an agent is a special type of autonomous controller that is, among other characteristics,

- multidimensional,
- multivariate,

- composite,
- variable, and
- metamorphic.

Animals often manifest some if not all of these characteristics. They also manifest coordination at many different levels of their organization, and generally manifest both weak and strong goal-directedness. In the following chapter, I will utilize the concepts and terminology introduced here to develop an account of the basic nature of agency and discuss the ways in which agential concepts non-metaphorically apply to robotic and organic control systems.

Chapter 5
Agents as Control Systems:
An Interdisciplinary Synthesis

> [W]e must be careful that the ways in which we construe agency and define its nature do not conceal a parochial bias, which causes us to neglect the extent to which the concept of human action is no more than a special case of another concept whose range is much wider.
>
> —Harry Frankfurt (1978, p. 162)

5.1 Introduction

In this chapter I will offer an account on which an agent is a special kind of control system: one that can change its perspective on whether or not something is under its control. This is the basic requirement for a system to have preferences; the possession of preferences is necessary and sufficient for something's being an agent. In Section 5.2, I will argue that agency is control by means of preferences, and explain what this implies. In Section 5.3, I explain how beliefs, desires, and intentions can be understood given the account laid out in Section 5.2. In Section 5.4, I briefly summarize my characterization of what an agent is, borrowing concepts from Grush and Springle's (forthcoming) discussion of the relation between agency and skills. Section 5.5 then situates the present discussion in the larger topic of "agent architecture" and addresses several terminological issues. Finally, Section 5.6 reviews the differences between the account of agency offered here and other attempts to provide minimal conditions for agency.

In this chapter I am offering necessary and sufficient conditions for agency. I am writing this as a philosopher but it is intended for an interdisciplinary audience. Some usages in some

domains (e.g., artificial intelligence and agent-based modelling) will often not meet all of my criteria. My intention here is not to legislate about who should get to use the word ‘agent’, and in what circumstances (this would, of course, be counterproductive and a fool’s errand). Instead, it is to define a standard notion of agency that can most effectively bridge across domains, and to offer the tools for understanding how different usages differ from and relate to one another, and how agential terminologies can be usefully translated. I am not offering the present account as the once and for all correct way to understand agency, but instead as a reference against which other notions can be usefully distinguished. More than merely offering a definition of agency, the framework presented here builds on the previous chapter to elaborate the kinds of features that control systems can have that make them similar to or comparable to “agents” in various ways. This can provide a coherent conceptual repertoire for qualifying usages of ‘agent’ and other agent-related terminology so that workers in different domains can exchange ideas without having to come to an agreement on who gets to own unqualified usage of the word ‘agent’.

5.2 Preference-Based Control as Necessary and Sufficient for Agency

Agency is a foundational concept in

- law (List & Pettit, 2011),
- anthropology (Rapport & Overing, 2000, pp. 1–9),
- economics (Ross, 2018),
- linguistics (Palmer, 2007, p. 1048),
- psychology (Bandura, 2001; Seligman et al., 2013),
- sociology (Emirbayer & Mische, 1998),

- neuroscience (Shadlen & Kiani, 2013),
- ethology (McFarland & Bösser, 1993),
- artificial intelligence (Russell & Norvig, 2010),
- robotics (Murphy, 2000, p. 70), and
- philosophy of action (Taylor, 1966).

What these fields arguably have in common is that they are focused on entities or systems that confront situations in which there are multiple ways they can respond and can *choose* how they will respond. This might sound, especially given the previous chapter, a lot like a description of control systems: a control system categorizes its situation and selects a response. But note the word *select*: a selection is not necessarily a choice. Sober's (1984, p. 99; see Figure 5.1) "selection machine" can select which balls will be distributed into the lower compartment, but it does not have the power of choice; it is not an agent. What differentiates an agent from other types of control systems, then, is the power of *choice*.

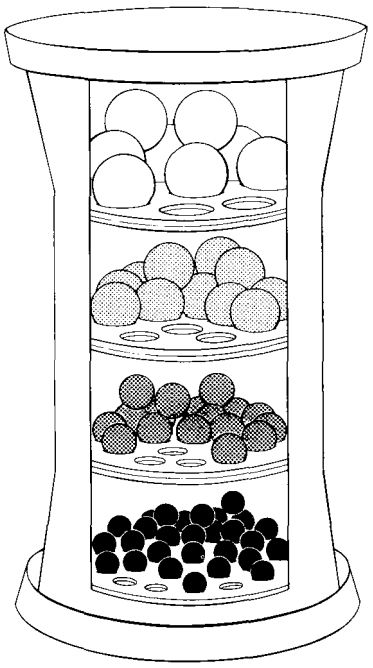

Figure 5.1: "Selection machine" from Sober (1984, p. 99).

What is the difference between the ability to choose and the ability to select? T. F. Daveney (1964) tackled this question directly, and argued that there are two basic differences between choosing and selecting:

- What gets chosen in any instance of choosing is a *state of affairs*. Other kinds of things, like actions or objects, are only picked or selected, not chosen.
- Choice always involves a valuation: the option that is chosen by the agent is the one that is considered best by the agent from among the available options.

Neither of these points is obvious or self-evident, and one or both of them might initially strike the reader as wrong. However, I think that these points do capture the difference between choosing and selecting, and are therefore key to understanding the basic nature of agency. For this reason, they are both worth examining closely.

5.2.1 What Are Chosen: States of Affairs

At first glance, this might seem like an obvious point. One might respond that grammatically, other categories can be converted into states of affairs. Rather than selecting the “on” setting of the furnace, we might say that the thermostat selects the state of affairs in which the furnace is turned on. Rather than selecting forward motion at a given moment, we might say that the robots selects the state of affairs in which it is moving forward.

However, this response misses something important. Representation of something *as* an action or *as* a control parameter or *as* an affordance for action is very different from

representation of something *as* a state of affairs.[72] The difference here corresponds to Grush's way of distinguishing between 'subjective' and 'objective':

> On one common use of these terms, they mark a contrast between things that are dependent on being represented, or being experienced, (subjective) on the one hand, and things that are truly independent of all representation: something that exists, or is true, regardless of the experience or biases of the representer. This is a perfectly fine usage, but it is not quite the contrast I am after. The contrast I am interested in is between things that are represented as being dependent upon the representer (subjective), and things that are represented as being independent of the representer (objective). In my sense of objective, then, the large toothy spiders crawling up the LSD user's arm are objective, so long as the hallucinator thinks they are real. The hallucinator has the tools that the [proverbial ostrich that seeks to deprive the predator of existence by hiding its head in the sand] lacks, he is just misapplying them. I am interested in understanding the tools, not in the conditions of their correct application. (2000, pp. 60–61; cf. Weiss, 1938, p. 29)

What is meant by "represented as being dependent upon the representer"? Suppose I represent something, such as a banana, solely as just an affordance for my own nourishment. Its identity *qua* affordance for nourishment is dependent on whether it can in fact nourish me and nothing else. If that is the only way I can register its existence, then I would have no ability to take into account the possibility of its being eaten by someone else, or the possibility of using it for some other purpose, of doing something with it besides eating it (like placing it somewhere to store it for later), or the danger of slipping on it and falling.

I could form separate representations for these purposes. I could recognize an affordance for slipping and falling, an affordance for nourishing someone else, etc. But it requires a yet separate feat for me to treat these as representations of the *same* object: to

---

[72] Gibson, for example, did not consider affordances to be objective features of the environment, writing that

> an affordance is neither an objective property nor a subjective property; or it is both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is equally a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer. (1979, p. 129)

represent that there is an object that can serve as an affordance for throwing a projectile at an enemy or for storing it for later nourishment, or even to remember that there is still the very same object in the same location, even though it is not currently an opportunity for throwing or nourishment.[73]

Daveney argues that when we make choices, we "do not choose objects as such, but rather states of affairs in which objects feature" (1964, p. 85). He offers the following example:

> Suppose I pick a certain flower in a meadow, and announce that I have chosen it. A friend remarks, "You are going to write a sonnet about it, or take it home to decorate a room?" I reply, "No. I haven't chosen the flower for anything—there is nothing I am going to do with it. I have just chosen it, for it is possible to choose things, isn't it?" And I steadfastly refuse to agree that there is any state of affairs relating to the flower which I commit myself to in making the choice. My claim to have chosen it, in these circumstances, is meaningless… (ibid., pp. 85–86)

Daveney does not use Grush's terminology, but his argument essentially is that choosing involves making a selection of an objective state of affairs against other objective states of affairs. It involves selecting, not between actions or control parameters or affordances or sensory or perceptual states (which are all *subjective* in Grush's sense), and not between objects (considered in isolation), but between states of the world. I don't think that this requires having a single "world model" within which all of one's representations are coordinated or semantically related, but it does require the use of representations or models that are objective in a certain sense, that is, that represent their contents as being independent of the representer. But *independent* in what sense, exactly? I tackle this question next.

[73] Treating multiple representations as referring to the same object actually involves two separate feats: there is the challenge of treating multiple affordances as a unified object (Cussins, 1992, p. 659), and there is the challenge of treating spatiotemporally and qualitatively disjoint representations as referring to the same object (or "reidentification"; Strawson, 1959, pp. 31–38).

5.2.2 A Perspective on What Might or Might Not Be Under My Control

My contention is that objective representation, i.e., representing states of affairs as states of affairs, involves having a perspective on something as being potentially under my control sometimes and not under my control at other times. Consider a simple example of how something can be represented as under my direct control in some ways but not others. If I move my head up, the cylindrical beverage can in front of me appears differently. The top looks rounder, less flat. But my mental registration of its three-dimensional shape doesn't change. I have not changed its shape or its position by simply moving my head up. I could alternatively reach out my hand and reposition the can to provide a similar sensory profile as if I had moved my head in the same way; now I represent the can as having been moved by me.

The can's shape and position exist apart from my sensory manifold and apart from my motor capabilities, and I represent the can and its shape and position *as* being separate from them. I can represent them as remaining constant even when my sensory input is changing, and when I am selecting different control parameter values (i.e., sending different motor commands) for my muscles. This is very different from simpler control systems, like the thermostat. The thermostat does not need to represent the *status of* whether the ambient temperature or the furnace setting is under its control in a given instance. As a matter of fact, the ambient temperature is not under its direct control, whereas the furnace state *is* under its direct control—but the thermostat doesn't represent *these* facts. The thermostat does not represent these as states of affairs that may or may not be under its control at different times.

Now suppose I form a preference: at the moment, I would prefer the can to be at a higher location (for example, on the shelf in front of me, instead of on the desktop) to reduce the risk of spilling its contents. I cannot control the position of the can unless I reach out my hand and grasp it first. However, if I move my head downward, my sensory input will change in a way that would correspond to the can moving upward if I were not currently representing the

status of whether its position is currently under my control (and in particular, whether it can be controlled by means of my head movements). My preference is not about the sensory input itself, or about the movement to be performed; it is about the position of the can, an objective state of affairs. In order to have such a preference about an objective state of affairs, I must keep track of whether and in what ways my movements can result in a change of this state of affairs.

Agents, therefore, unlike other control systems, can implicitly regard something as being controllable or uncontrollable in different ways at different times. The reason why this is key to agency is because it is the point where the semantic content of the system's internal states becomes decoupled from the direction of fit.[74] A non-agential control system has sensory states that have a certain content, and can make behavioral selections that have a certain content. The sensory contents must *always* be sensory contents, and the behavioral contents must *always* be behavioral contents. The sensory contents always have a world-to-system direction of fit, and the behavioral contents must always have a system-to-world direction of fit.[75] It is only when the semantic contents become decouplable from the direction of fit that behavioral explanations become intelligible in terms of the system's own *reasons* (rather than in terms of an external observer's "intentional stance"), which as Anscombe (1963, pp. 7–9) and Dretske (1999, pp. 19–20) have argued is essential to agency. Another way to put this is that there is a single semantic repertoire that is common to the control system's input and output states.[76] Reasons explanations require a space of reasons (McDowell, 1994, p.

[74] The phrase 'direction of fit' was coined by Searle (1979) but the distinction between world-to-mind and mind-to-world direction of fit was explicated in different terminology much earlier by Montague (1909, pp. 234–235).

[75] Millikan (1995) argued that simpler systems than agents can also have what she calls "pushmi-pullyu" representations, which have both directions of fit at once, or as she puts it, both a "descriptive" and a "directive" function. See also Chemero (2009, p. 26).

[76] Though reasons explanations are often thought to require propositional mental attitudes, I instead agree with Heck (2000) that a system's reasons may be nonconceptual. On the other hand, the present account would fit well with the situation semantics developed in Devlin (1991). Devlin takes types of

125), a repertoire of representational contents that enable control system states to be directed at states of affairs in the manner of a belief or a desire.[77]

I consider this way of understanding the *independence* involved in objectivity as superior to that of Cussins, who writes that "An objective world is given to a subject if the content presents something as being independent of the subject's particular abilities, and particular location in space and time," (1992, p. 659). The content need not be totally independent of the subject's abilities, and I would argue (though I won't do so here[78]) that it *cannot* be so independent. Objectivity need be neither a "view from nowhere" (Nagel, 1986) nor a "view from anywhere" (Cussins, 1990, p. 426) nor a "God's-eye" point of view (Evans, 1982, p. 152), and it need not be domain-general (Premack, 2007). It merely must be representable by the agent (or "subject") as both potentially controllable and potentially uncontrollable by that agent. This is a much thinner notion of objectivity that need not satisfy Evans's (1982, p. 104) generality constraint and that does not require any kind of robust practical rationality.[79] But I believe it captures the minimal core essence of what separates agents from other control systems.

One question that needs to be tackled is: *how* can a control system have a variable perspective on whether something is under its control? We will be in a better position to answer this question after we consider Daveney's second difference between selecting and choosing.

---

situations or states of affairs, and the constraints that hold between them, to be the basic semantic building blocks, and any system that works with a determinate scheme for sorting situations into types will thereby possess a "concept" on his account (1991, p. 19). This is of course a much thinner way of understanding 'concept' than most participants in debates about "conceptual/nonconceptual" content (such as Heck) have in mind.

[77] Here my view is in disagreement with Hurley, who writes that "the space of reasons is the space of action" (2003, p. 231); it is instead the space of representations of states of affairs that can be the contents of beliefs or preferences.

[78] Such a defense would require a fully spelled-out, thorough-goingly embodied/situated/embedded theory of semantic content, which is unfortunately out of scope for the present dissertation. I believe that the present dissertation could potentially provide the foundation for such a theory, however.

[79] Though Davidson points out that "there is a limit to how inconsistent a person can be and still be credited with clearly defined attitudes" (1999, p. 8).

5.2.3 Selecting Based on Which Option is Considered to Be the Best

Daveney argued that when making a choice, the option that is chosen by the agent is the one that is considered best by the agent from among the available options. To see this, Daveney points out that there are other instances in which agents make selections that we would not call choices: "I cannot be said to choose, if I am merely looking for an object answering a certain description among a lot of other things" (1964, p. 84). In such a case, it will sound strange to call it "choosing" if there is no dimension or criterion of *goodness* involved in the description. Daveney provides the following example:

> To illustrate this distinction, suppose I am about to perform a certain operation in carpentry. The job can, I know, be done with any one of a number of tools, although not all the tools will do the job equally well. I pick the one that I think will do the job better than anything else, and my decision to use this one as the most suitable, and my rejection of all the others, is a genuine case of choosing. This is a different case from the one in which I merely look for a pair of long-nosed pliers among a heap of pliers on the bench. In choice I am guided by an end in view, and what is chosen is what I think will best bring about this end. I evaluate the alternatives in the light of my aim and select the best. (1964, p. 84)

Daveney considers another example in which he is asked by someone performing a card trick to "pick out a card" from a number of cards that are face down. Assuming that there is no prior aim or criterion of goodness that the selection satisfies, Daveney argues that "It cannot be said that I chose carefully or carelessly, etc., and I cannot answer the question, 'What did you choose that one for?'" (1964, p. 84). Similarly, McCall writes: "selecting moves in a random way could hardly be called choosing" (1987, p. 285). What, then, does it mean for there to be a criterion of goodness that guides the selection? My contention here is that there must be an antecedent *preference*: when *choosing* state of affairs A over state of affairs B, the agent must *prefer* A to B.

This might seem like a trivial point. In fact, according to the way economists sometimes define 'preference' (i.e., "revealed preferences"), the existence of a preference is *constituted by* the fact that an agent chose A over B. This will not be my usage, and I will also not consider

preference to merely be a *disposition* to choose one state of affairs over another (Weirich, 2013 refers to this as a "behaviorist account of preference"). A preference is an occurrent feature of a control system; it is what Dretske (1988) called a "structuring cause." In the card trick case, Daveney picked a card but not based on any antecedent preference.

More specifically, a preference is a certain kind of modification to a control system's input/output mapping. It is like an implicit rule (in the Sellarsian sense of "rule-governed behavior"; see Section 3.5.2) that is operative in a control system that says that when presented with certain alternatives, one of them should be selected instead of the others (where 'should' is being used in the same razor-thin sense in which the alpha creatures from Chapter 3 operate with an implicit rule that beta creatures "should" be censured under certain conditions). But it is an *effective* rule: while it has the preference, the preference is an operative part of the input/output mapping of the control system. Further, the alternatives to be selected between are states of affairs. Weirich writes that "Preference is a mental state that compares two situations. To prefer one situation to another is to favor the first situation; one would rather be in the first situation," (2013, p. 4041). Possession of this implicit rule requires that the control system must be able to *recognize* the presence of an alternative of states of affairs that can be selected between. What exactly does this mean?

5.2.4 Recognizing Alternative Outcomes

Sterelny (2003) distinguishes between what he calls "drive-based" and "preference-based" motivation. Essentially, drive-based motivation occurs when a response is selected based on what kind of situation is detected as being present. What is selected in this case is a *behavior*, not a situation (or state of affairs). This involves the control system's possession of

some kind of functional mapping from detected situations to behaviors.[80] In this case, the kind of thing that the control system selects is not the same kind of thing that it detects. It does not need to register or represent what kind of situation will *result from* the behavioral selection, and it does not select behaviors on the basis of what kind of situation or state of affairs will result. Since it is not states of affairs that are *selected by* the system, it cannot be said to operate on the basis of preferences.

Sterelny, following Anthony Dickinson (e.g., Dickinson and Balleine, 2000), argues that we can infer the presence of preference-based motivation when an organism can implicitly take into account the causal consequences of their behaviors when selecting those behaviors. They do not blindly favor the same behaviors in the same circumstances if it becomes apparent that those behaviors will lead to states of affairs that are less desirable. In this case, the state of affairs itself is evaluated, and behaviors are selected on that basis. In order for behaviors like these to be learned and reinforced, the organism must be capable of detecting the presence of some state of affairs S and of anticipating the fact that a certain behavior B1 will causally produce some state of affairs R1, and an alternate behavior (or lack of behavior) B2 will result in some state of affairs R2 (S might be identical to or represented as identical to R1 or R2, but R1 must be represented as distinct from R2 for one to be preferred, of course). It is then R1 and R2 that are directly compared, not B1 and B2. Either B1 or B2 is selected based on whether R1 or R2 is preferred.

[80] In such contexts, the term ‘motivation’ is sometimes used for relatively complex organisms, robots, or artificial agents. It is difficult to discern any kind of rule dictating when a control system is complex enough to count as incorporating “motivation”. In this dissertation, I will not take a stand on this issue. One definition that is often useful is the following: M is a motivational variable for a control system if it is an operational part of the mapping from input states to behaviors, and does not vary in any simple way in relation to any input or output states of the system. This definition becomes less useful in contexts that rely on a distinction between “conative” and “cognitive” states, aspects, or activities. In the latter cases, the definition of what counts as “motivation” usually depends on the kind of agent architecture under discussion; see Section 5.5.

The organism must be able to:

- Detect whether R1 obtains or not, and whether R2 obtains or not
- Determine what behaviors can currently be performed
- Anticipate what states of affairs will result from performing behaviors given the states of affairs that currently obtain[81]
- Anticipate whether R1 or R2 will be among the results of actions that can be performed given the current states of affairs

Sterelny's position is compatible with Irwin's earlier characterization of choice and preference: "A choice, then, expresses a preference for one of a pair of differential outcomes over the other, in some common outcome field" (1971, p. 7). One study that Dickinson and Balleine performed to determine whether rats possessed these capacities involved what they refer to as *devaluation*:

> hungry rats are trained to press a lever to receive food pellets. These pellets are then devalued by conditioning a food aversion to them. This aversion is conditioned by allowing the animals to eat some of the pellets immediately before they are made to feel sick by an injection of lithium chloride. After a few experiences of this association between the pellets and sickness, the animals learn to refrain from eating the pellets. As this aversion conditioning takes place in the absence of the opportunity to press the lever, the animals cannot associate this action directly with the illness. Even so, they are reluctant to press the lever when once again given the opportunity to do so. This test is conducted in the absence of any food pellets so that the low frequency of pressing shows that the performance of this action is based on knowledge of the action-pellet relation acquired during initial training. (1995, p. 162)

The rats must therefore be capable of selecting actions, not due to an antecedent preference about the actions themselves, but due to an antecedent preference about their outcomes. The rats detected that the sickness state of affairs accompanied the ingestion of food pellets. In doing so, the sickness was registered as a perceptual state, not a behavior. But the rats also treated this same situation as one that they could control in the future; they could make

[81] MacCorquodale and Meehl (1954) discuss the minimal kind of "cognitive map" that would be needed for this capability, which (following Tolman) they refer to as "expectancy."

behavioral selections in the future to avoid that same state of affairs. They are able to have mental states with either a mind-to-world or world-to-mind direction of fit with the same contents.

5.2.5 Selecting Based on Which Option is Considered to Be the Best, Reconsidered

At this point, Daveney might not be satisfied that the rats in Dickinson and Balleine's experiments count as genuine choosing, for reasons that Sterelny articulates. Consider another type of experiment. Dickinson and Balleine (2000) trained rats to engage in separate behaviors to receive either carbohydrates or proteins. However, the rats are always kept hungry, and never allowed to experience being satiated with respect to either carbohydrates or protein. They are then prompted with a novel situation: they are allowed to become satiated with respect to one of the food sources. After this occurs, the rats will choose only the behavior that leads to the other food source. Dickinson and Balleine argue that this shows that the rats are driven by preference for outcomes; they cannot have learned such behaviors through conditioning because they had never engaged in the behaviors in such circumstances before.

Sterelny (2003) argues that though this kind of result is consistent with outcome-based preferences, it might also be that they possess separate motivational systems that keep track of action-outcome contingencies, but at the same time, which motivational system wins out in a given case might be a matter of its drive strength, rather than an outcome-based utility comparison.[82] For this reason, Sterelny argues that the possession of preferences requires that competition not just between candidate behaviors but also between disparate motivational systems must be outcome-based. If motivational systems compete for behavioral selection

[82] The notion of a "motivational system" was introduced in Section 4.6.3.4. For more recent arguments based on brain and behavioral research that action-event contingencies are used across different motivational systems, see Hommel and Wiers (2017). For a recent review of investigations into outcome-based motivation in animals, see O'Doherty, Cockburn, and Pauli (2017).

based merely on drive intensity or hard-wired hierarchical relationships, such behavioral selections would not count as being based on comparison and valuation of outcomes.

The ability to choose behavioral candidates generated by competing motivational systems prompts the question of what is required to make this possible. In the next section, I argue that the rudiments of beliefs, desires, and intentions emerge for systems capable of making outcome-based behavioral selections across the domains of different motivational systems.

5.3 Beliefs, Desires, and Intentions

5.3.1 Likings

The idea of preference-based outcome selection might suggest that an agent must possess a database in which every possible outcome pair is represented along with the outcome that will be preferred when presented with that pair. Pollock (2006, Chapter 2) provides good reasons to think that this would not be computationally feasible. It would be far more economical for any preference-based agent to store objective representations of cardinal value to types of situations, and then generate binary preferences on the fly when presented with a choice by comparing the cardinal values. Pollock argues that biologically such cardinal values are probably analog representations, which are measured when a choice must be made. Such value representations, on his view, should not be thought of as preferences or desires, but instead as "feature likings"; they are not desires because, being likings, they do not represent an active motivational energizing of behavior.[83,84] A desire, properly speaking,

[83] I elaborate on the notion of "energizing" behavior in Section 5.3.3.
[84] Pollock's distinction between liking and desire should not be confused with Berridge's (2018) distinction between liking and wanting. By 'liking', Berridge refers to the hedonic component of reward. Pollock's notion of 'liking' is not hedonic, but is instead more closely connected to what Berridge calls 'wanting'.

implies that the agent is in an active state of seeking a certain outcome but has not necessarily adopted behaviors (or a “plan”) to acquire the outcome (the latter would imply an *intention*, not merely a desire).

An agent similarly does not need to store a comprehensive database of feature likings to generate preferences about outcomes on the fly. Pollock argues that “The only way to make evaluative cognition work is to have a database of ‘computationally primitive values’ from which other values not included in the database can be computed” (2006, p. 35). Here it is important not to take computational metaphors too literally; Pollack is not advocating a Fodor-style computational theory of mind but merely making a point about basic features of the information processing that is necessary. However, it is the need to work out derivative feature likings from the “computationally primitive values” that makes beliefs necessary. This is because systems capable of beliefs can represent causal relationships between states of affairs in a way that abstracts away both from particular motivational or evaluative systems, and from what the system is currently able to control, as I explain next.

5.3.2 Beliefs

A number of philosophers have debated about what the capacity to believe minimally requires. Davidson (1982) argued that belief possession requires possession of language; Stich (1978) argued that belief possession requires the possession of concepts. Philosophers sometimes argue on either of these bases that organisms other than humans cannot have beliefs. Such philosophers may well be right that language or concepts are necessary for possession of certain *kinds* of beliefs. In this discussion, I do not wish to take a stance on what counts as a language or what counts as a concept, or which kinds of states should or should not be counted as “beliefs”. However, I have been articulating an account of agency that is based on analyzing the concept of a *choice*, and on the assumption that choices are unlike

mere selections by being based on *preferences*. The notions of choice and preference that I have been working with are grounded in the account of control I gave in the previous chapter, which has so far been noncommittal with respect to linguistic or conceptual capacities. I believe that this route to an understanding of agency leads to a corresponding way of understanding *belief* that is similarly neutral with respect to the possession of linguistic or conceptual capacities by the agent (i.e., both to whether the agent must possess them and what it means to possess them).

Sterelny's (2003) distinction between drive-based and preference-based motivation comes packaged with another distinction that is important to his account, between detection-based sensory tracking and what he called "robust" tracking. Detection-based tracking occurs when an organism relies solely on a single cue to track any given feature of its environment. Organisms that rely on this type of tracking can be easily fooled if there are other features in the environment that can cause false positive registrations for the cue. Instead, robust tracking capacities are often selected for in lineages that have to cope with greater environmental complexity. With robust tracking, an organism is able to integrate information from the detection of multiple cues to keep track of a single feature.

Sterelny argues that robust tracking is not yet sufficient for belief, however. Robust tracking can occur in conjunction with drive-based motivation: behaviors may be driven directly by the registration of features that are robustly tracked, without the mediation of outcome valuations. For Sterelny, belief emerges when such registrations count as "decoupled representations," that is, "internal states that track aspects of our world, but which do not have the function of controlling particular behaviors" (2003, p. 29). He therefore defines beliefs as "representations that are relevant to many behaviors, but do not have the biological function of directing any specific behavior" (ibid.). Sterelny (2003, p. 86) notes that preference-based motivation may not be necessary for the possession of beliefs in this sense (since the function

of such beliefs in a given organism may simply be to modulate a wide range of drives or modulate drive-based behaviors in diverse ways).

Above, I argued that representing states of affairs *as* states of affairs, which is necessary for choosing and having preferences, involves having a perspective on something as being potentially under my control sometimes and not under my control at other times. In this way, a content is not merely decoupled from particular behaviors; it is decoupled from *control* as such, *for the agent*.[85] I believe that this makes for a more objective (i.e., external observer-independent) criterion for what counts as a belief. To have a belief, then, involves more than just the registration of information that is not treated as intrinsically relevant to any behavior in particular. It must have a content that could also be the content of a preference *for that agent*. If I do not have the *capacity* to prefer that the sky be blue, then I cannot have the *capacity* to believe that the sky is blue. This may strike the reader as wrong on first glance: we usually only form preferences about things that we consider to be controllable. This is true; but it only shows that we do not *actually* form preferences about things that we don't consider to be controllable. If I were presented with a set of buttons, and I found that pressing different buttons changed the color of the sky, I would be capable of forming a preference about which button to press, i.e., what color to make the sky. Being able to envision or anticipate a means of controlling a situation is not necessary; what is necessary is that *if confronted with a means of controlling* the situation, one could form a preference and make a choice. My contention is that *this* is exactly what it means for something to be an agent, rather than another type of control system.

If beliefs must have content that could also be the content of a preference for the agent, then this shows how systems become capable of making outcome-based behavioral selections across the domains of different motivational systems. On the sensory side, perceptual contents

[85] For an account of how a common repertoire of contents for sensory and action representations may be implemented in the brain, see Hommel & Wiers (2017).

are unified into a common spatial field in a way that motivational systems often do not unify behaviors into a common behavioral space. By learning new correspondences between states-of-affairs *represented as* achievable outcomes for particular motivational systems, on the one hand, and states-of-affairs *represented as* perceptually distinct situations, on the other, a content framework that spans across motivational system domains that is also neutral with respect to direct-of-fit (i.e., decoupled from the status of being-under-control or not) becomes possible. This way of understanding *belief* is another piece of the puzzle that enables us to understand agency (and related notions) in a way that distinguishes agents from other control systems but also places at center stage the key features that a wide swath of disciplines is unified in considering to be important about agency.

5.3.3 Desires and Intentions: Agential Goal-Directedness

Providing a bare sketch[86] of what desires and intentions look like on this account requires elaborating on my above usage of the phrase "active motivational energizing of behavior" when discussing the difference between liking and desiring. States of *liking* are often not considered motivational or conative because merely liking something does not in itself involve *energization* of behavior, as it is often put by psychologists (e.g., Schultheiss et al., 2012). As one philosopher wrote,

> a reason or motive is a moving or impelling thought, the thought of that for the sake, or in view of which, some act is done; and I myself see no intelligible alternative to saying that it "moves" or "impels" in the sense that it functions as a cause of actions, in the conventional sense of cause as an antecedent implying a consequent by a rule of invariable connection. I should therefore, describe a motive as a *causa rationis*, a mental antecedent which, when attended to by a person, and in otherwise comparable conditions, will invariably be followed by an orientation of his organism towards the action thought of, in a way which, except for the intervention of distractions, counter-motives and physical impediments,

[86] Both the nature of desire and the nature of intention (and for that matter, belief) are huge topics in themselves that have received multiple book-length treatments, so more than a minimal sketch will not be possible here, though a little more substance will be added in the next chapter.

> will terminate in the action itself. Such a thought may be said to constitute simply a reason or motive, if, when attended to by itself alone, it thus causally implies action... (Falk, 1948, p. 116)

Merely liking a state of affairs may be said to "causally imply" a preference (or preferences), perhaps, but it does not "causally imply" action. A preference only implies an action if one is in a situation where one must make a choice. This may make one desire the preferred option (if only out of a desire to escape the alternative options), but the desire to avoid the choice situation itself may easily prevail (and it may easily lead to an extinguishment of the desire for the preferred option) if a means of doing so becomes available. As Schroeder puts it, "the ability to get the body moving," sometimes also called as "oomph" by philosophers (2004b, p. 22), is essential to desire.

Another way of making the same point is that desiring, unlike liking, initiates and sustains striving for a state of affairs. This way of putting it makes salient the goal-directed nature of desiring as opposed to liking. As I discussed in the previous chapter, "goal-directedness" does not refer to any one phenomenon but instead a range of distinct kinds of phenomena. The specific kind of goal-directedness that is key to understanding the difference between liking and desiring is what I referred to as "weak goal-directedness" (Section 4.6.3.3). Weak goal-directedness at minimum requires a *variable* control system that can change its input/output mapping over time (Section 4.6.3.2). It involves what I referred to as *targeting*, which is the online *coordination* (Section 4.6.2) of input parameters with behavior parameters. Desire requires not merely tracking whether the desired state of affairs (i.e., *goal*) obtains, but also progress toward the obtaining of that goal (the input parameters); behavior is then actively coordinated with such progress registrations. This might make it sound like desire-driven behavior is limited to simple sensorimotor coordination, but as discussed in Chapter 4, such progress tracking may involve higher-order "quasi-spatial" (in the terminology of Grush, 2000) manifolds, or "order parameters" (Kelso, 1995). This is true even for the most sophisticated

desire-based behaviors such as pursuing long-term career goals or spiritual goals (e.g., seeking "enlightenment").[87]

Several things set desire-driven behavior apart from mere weak goal-directedness: 1) the fact that the goal is represented *as* a state of affairs, 2) the use of *beliefs* in tracking progress toward the goal (the inputs for the coordinating control subsystem are parametrized in terms of beliefs), and 3) parameterization of the outputs of the coordinating control subsystem in terms of intentions. The adoption of a desire is the reconfiguration of the input/output mapping of the control system such that beliefs are coordinated in a certain way with intentions: intentions are continually updated in accordance with the updating of beliefs.

What, then, are intentions? Intentions[88] are to motor commands what beliefs are to sensory detections. Intentions map representations of states of affairs (the *contents* of intentions) ultimately to basic behavioral *skills*. A basic skill is a control subsystem (a motor schema, or a higher order behavioral schema; Arkin, 1998, p. 43) that generates motor commands without the mediation of the agent's beliefs and preferences. We can give determinate meaning to this statement in light of what has come before: The details of how behavior is directed by basic skills are not, according to the agent's own perspective, controllable (at least, controllable *during* the behavior) by the agent.[89]

For example, suppose I form an intention to lift a beverage can to a higher location. For most people, such a movement can be performed without having to reason about means and ends, i.e., without having to form the intention to lift my arm, open my hand, grasp the can, etc.

---

[87] Though such long-term goals will often involve committal, rather than inclinational, desires, and therefore a more complex type of coordination than has been discussed so far; see Chapter 6.

[88] Here, my focus is on what Searle (1983) calls "intentions in action," that is, intensions that direct occurrent behavior, rather than "prior intentions," planned actions that are not yet under execution. Pacherie (2008) refers to intentions in action as "proximal intentions" or "P-intentions"; importantly, these are distinct from what she calls "motor intentions" or "M-intentions." By 'intention', in this chapter, I will only mean proximal intentions, not motor intentions, which are internal to basic skill systems and generally not directly coordinated with beliefs and desires. What Searle refers to as "prior intentions" and Pacherie refers to as "distal intentions" will be discussed in the next chapter.

[89] Polányi (1958, p. 56) distinguished between intentions and the behavioral skills that underlie them in a similar way.

For a normal adult, lifting the beverage can to a higher location is a task that can be performed automatically by relying on motor skills alone, rather than practical reasoning. Similarly, one can act on one's intention to continue riding a bicycle in a particular direction without needing mediating intentions or preferences about whether the handle bars should be turned slightly to the left or right at a given moment to maintain balance. Objective representations are only necessary for forming the initial intention; one's motor skills then take over.

5.4 What is the Agent?

The agent, then, is a multidimensional and multivariate control system (capable of tracking many sensory and behavioral parameters) which is also necessarily composite. Agents consist of many different control systems combined into a central control system. The central controller has beliefs and intentions in its input and output repertoire, but the contents of beliefs and intentions form a common repertoire of contentful states that represent states of affairs. If this central control system can change its repertoire of beliefs and intentions over time (e.g., by learning and forgetting), then it is a metamorphic control system. Its input/output mapping will necessarily incorporate both preferences and desires (and likely, if Pollack is right, other states like likings).

The central controller will have other controllers whose job it is to update beliefs based on sensory input (these implement *perceptual skills*) and controllers whose job it is to translate intentions into motor commands (these implement *behavioral skills*). Grush and Springle (forthcoming) characterize these as controllers that implement *inverse mappings* because in the case of perception, they determine what kind of state of affairs would causally produce the sensory input, and on the behavior side, they determine the motor commands based on what kind of motor commands would causally result in the intended state of affairs. The possession of a central belief/desire controller is necessary for agency but the agent itself should be

identified with the larger containing control system (Taylor, 1966, pp. 134–138). The central control system is the larger agent's *means* of having a perspective on what is under its control and what is not.[90] The junction between the central controller and its skill controllers corresponds to what Grush and Springle refer to as an "inverse intersection" (forthcoming, p. 9) because it is the intersection of the perceptual inverse mappings and behavioral inverse mappings. The interface between these mappings and the central controller determine what Grush and Springle refer to as the "subjective accusatives of agency" (or SAAs), which define the limit of what the agent implicitly takes itself to be able to directly perceive and directly control. In this way, as Grush and Springle write, "Control, from a subject's point of view, starts with an SAA. And SAAs are defined by inverse mappings," (forthcoming, p. 6).

5.5 Agent Architectures and Terminological Issues

The discussion so far has omitted an enormous amount of detail, on questions such as:

- How are preferences formed? Does this require use of a "common currency" (McFarland & Sibley, 1975) representation of utility/reward?
- What other kinds of states (likings, goals, emotions, hedonic states, affective states, neuromodulatory systems, internal "need" states, etc.) are functionally implicated in the formation of preferences, and how does this work?
- How are beliefs used to update beliefs, preferences, desires, and intentions? Are there alternatives to reward-based learning? To what extent are "rational" processes of judgment and inference versus heuristics used? To what extent are these processes "penetrable" by peripheral control systems?

[90] I will discuss reasons in the next chapter why the "central controller" (in the sense I am using here) may not straightforwardly map onto the "will," "consciousness," or "self" of the agent.

- How much behavior is controlled through the central controller (or "intentional") as opposed to peripheral (or "habitual") control systems? And how do the peripheral control systems work?
- Where does "cognition" begin and end in the complex arrangement of control systems? Which processes are "cognitive," and which are not? Which are "conative" or "motivational," and which are not?
- How are desires formed?
- How are intentions/goals formed? Can this occur in the absence of any top-level goals or intentions (Carbonelle, 1982; this is sometimes referred to as "goal autonomy", e.g., Castelfranchi, 1995)?
- What is the nature of the coordination process by which desires generate intentions?
- Can there be more than one central controller? Can they be nested within each other? Can they be hierarchically/heterarchically arranged? Can an agent have multiple central controllers functioning in series or parallel?

The present account does not attempt to answer these questions. This is not because the account is deficient or incomplete, but instead because they are questions, not about the basic nature of agency, but about different ways to *implement* agency. Any given agent will embody answers to these questions in the detail of its makeup, which AI researchers refer to as an *agent architecture* (Wooldridge & Jennings, 1995).[91]

Not all of what AI researchers refer to as agent architectures will yield a literal agent according to the present definition (i.e., a control system capable of *choosing* states of affairs, not merely selecting behaviors). For example, Braitenburg vehicles (Braitenburg, 1984) are

[91] This is not to say that theoretical questions about agential concepts are autonomous from or independent of implementation questions. I see them not as qualitatively discontinuous, but rather as occupying different positions on a smooth continuum of levels of grain.

sometimes considered to embody a very simple agent architecture, even though they operate by extremely simple dynamics (e.g., turning the wheels left instead of right when a blue light is detected). The robots of Rodney Brooks are sometimes referred to as embodying a “reactive agent architecture” (Wooldridge & Jennings, 1995, pp. 132–133), meaning that their behaviors are determined based only on their momentary sensory inputs.

Often, the interest of these “agent architectures” is not necessarily in understanding agency as such, but in the attempt to engineer systems that behave in agent-like ways, or reverse-engineer systems that behave in agent-like ways. This goal is not only important for robotics and artificial intelligence, since an enormous about of behavior and control processes within organisms (even humans) are by means of processes that would not count as fully “agential” on the present account. This is true to such a great extent that it would probably not, in general, be useful, even when adopting the present framework for understanding agency, to restrict usage of terms like ‘belief’, ‘desire’, ‘preference’, etc. to *only* those states that represent things *as* states of affairs.

Even when it is best to restrict language like ‘belief’, ‘preference’, and ‘choice’ to those systems that are *capable* of representing something *as* a state of affairs (for reasons I have discussed above, e.g., to have a notion of agency that is not grounded in external perspectives), in many contexts it is useful to adopt a way of talking on which even sub-agential motor commands involve “intentions” (such as in Pacherie’s use of ‘motor intention’) or basic sensory registrations count as “beliefs.” In fact, we often refer to behaviors performed by agents as “actions” when they are caused by habits or drives, or even when they are not even “goal-directed” in these ways (e.g., the action of accidentally knocking over a glass). And in other contexts, it may be useful simply to adopt a purely instrumental intentional stance for systems that would not be considered agents at all on this view, to say things like that “the bacterium kept going straight because it believed it was heading towards a food source.”

Such usages will, of course, befuddle social psychologists and economists. The power of the present framework is that the conversation does not have to end at this state of befuddlement: the present framework offers a means to express the full range of ideas, across contexts and even disciplines, while avoiding such ambiguity and confusion. However, a great potential source of confusion remains, because even in fully agential contexts, concepts like ‘belief’, ‘desire’, ‘preference’, and ‘intention’ are used in starkly different ways: Sometimes they are used in a way that implies a *commitment* on the part of an agent, and sometimes not. This is a large topic that cannot be addressed here and will instead be the subject of the next chapter. There, I will demonstrate that whether or not agential states involve commitment makes a huge difference to their functional profile. For example, in general, systems capable of genuine planning or future-directed intentions will likely require commitment-based states. In the rest of this chapter I will compare the present framework to other contemporary attempts at understanding agency.

5.6 Comparison to Other Accounts

5.6.1 Sterelny

Sterelny’s account of agency is in agreement with the one presented in this chapter in that it takes the capacity for preference-based behavior to be a necessary condition for agency (2003, pp. 32–33). On Sterelny’s view, both beliefs and preferences are attitudes directed at states of the world, and preferences motivate behavior if and to the extent that that behavior is expected to produce the preferred outcome. The difference between Sterelny’s view and the present one lies in how Sterelny spells this out.

As I explained above, what Sterelny calls “drive-based” motivation differs from “preference-based” motivation in that drive-based motivation, unlike preference-based

motivation, involves a functional mapping from detected situation kinds to kinds of behaviors or drive-specific goals, unmediated by expectations about the causal outcomes of those behaviors.[92] What is key for Sterelny is whether or not the system selects behaviors based on anticipating both the effect the behaviors will have on the world and whether such effects will be good or bad. Note that what is important is the fact that outcomes of actions (and valuations of such outcomes) are represented; this requirement, by itself, does not involve any kind of essential link between the contents of preferences and sensory contents. In fact, recall that for Sterelny (2003, p. 86), the capacity for preferences may conceivably exist in the absence of a capacity for belief. This is because the representational capacities necessary for preferences are not essentially tied to the representational capacities necessary for belief. Sterelny requires preferences and beliefs to be representational, but he does not require them to operate with a common repertoire of representational contents.

The problem here is that Sterelny's account does not actually offer a way of distinguishing between the anticipation of changes in states of the world, on the one hand, and anticipation merely of changes in (potentially complexly related) control parameters, on the other. Sterelny's account does not rule out cases in which the kind of thing that the control system selects is not the same kind of thing that it detects. Sterelny considers preferences to be representations directed at states of the world, but unless these include the coordination and registration of correspondence between sensory and control order parameters amounting to what Grush and Springle refer to as an "inverse intersection" (forthcoming, p. 9), such contents will not be represented *as* states of affairs or states of the world, but merely affordances for control with an indeterminate relation to the world.

What I have argued, by contrast, is that objective representation requires the ability to couple and decouple one's sensory contents and one's control parameter contents. It requires

[92] What Sterelny calls drive-based motivation corresponds to a type of what I called "strong goal-directedness" in the last chapter.

the ability to register one's sensory contents as *corresponding to* one's control parameter contents (by updating how these are coordinated in the central input/output mapping), and to alter these correspondences over time as one's sensory and behavioral situatedness changes with respect to the states of affairs they represent. This is necessary for states of affairs to be representable by the agent as both potentially controllable and potentially uncontrollable by that agent. To have a belief involves more than just the registration of information that is not treated as intrinsically relevant to any behavior in particular. It must have a content that could also be the content of a preference *for that agent*.

Sterelny seems to rely on the fact that external observers can observe the seemingly agential behavior of rats and *infer* that the rats must be representing behavioral outcomes *as outcomes*, i.e., as states of affairs. What I am offering is an account that relies exclusively on what I have called *internal* perspectivalism: representation that is grounded in the system's own perspective. By registering, maintaining, and updating correspondences between sensory contents and control parameters, agents identify states of affairs as things that can correspond to different kinds of sensory and control parameters at different times, and therefore things whose existence is not dependent on the system's sensory or control parameters. The main difference with Sterelny is that instead of taking for granted the capacity for representing states of affairs, my account explains what kinds of control system features this requires.

5.6.2 Burge

Burge (2009, 2010) develops a notion of agency that is much thinner than mine. He writes that

> Primitive agency forms a background for understanding both representation and representation-as in perceptual systems—hence for understanding perceptual kinds. Primitive organismic agency is phylogenetically prior to perception. It occurs in animals that demonstrably lack perception in the sense that I will elaborate. ... Usually discussion [of agency] begins with cases involving desire,

> intention, will, and then focuses on subcases of intentional action. There is nothing in itself wrong with this focus, of course. But often it is assumed that such approaches encompass all action. Animal action begins earlier. Much of it is pre-intentional, even pre-representational and pre-perceptual. Origins of agency precede those of perception and representation. Even *representational* agency precedes intention and belief, not to say meta-evaluation. (2010, p. 327)

According to Burge, not all biological movement, activity, or change counts as a manifestation of agency (i.e., as an "action"). What distinguishes actions from other activities is that actions, unlike other internally-produced changes in an organism, either "issue from," or are "a product of coordination with," at least some of the "central behavioral capacities" of the organism (Burge, 2009, pp. 262–264). Burge summarizes his view by writing that

> I think that the relevant notion of action is grounded in *functioning, coordinated behavior* by the *whole organism*, issuing from the individual's *central behavioral capacities*, not purely from sub-systems. Coordination is meant to imply that the behavior must issue from central capacities, in effect coordinating sub-systems, or coordinating central capacities with their peripheral realizations. (2009, p. 260)

Key to Burge's view is the distinction between what he calls "central capacities," on the one hand, and "sub-systems" or "peripheral systems," on the other. Burge never provides an explicit criterion for what counts as a "central" versus "peripheral" or "sub-" system, but instead provides a number of examples. For example, he writes that

> An animal's shivering in the cold, or its coughing or sneezing, are perhaps instances of behavior. But they are not instances of active behavior. The events can be functional. Shivering engenders heat. Coughing and sneezing have expectorant functions. They are functional, but they are operations of peripheral systems that are not normally products of coordination with central behavioral capacities. (Burge, 2009, pp. 261–262)

Burge does not, however, require that all actions must be exercises of biological proper functions:

> There are *types* of non-pathological primitive agency that do not obviously fulfill larger biological functions. Idly, non-intentionally, drumming one's fingers, or the unconscious coordinated swaying to rhythmic sound by an animal, can be active. It is not evident what function it performs. Certainly it need not realize any of the basic biological functions. (Burge, 2009, p. 262)

Further, even organisms without nervous systems possess central behavioral capacities:

> One can make a start at analogous points even for simple organisms like paramecia that lack a central nervous system. Eating involves a unitary process that involves the whole organism (eating itself, and rotation of the animal body so that the side that has the gullet opening faces the food), as well as operations that are imputable purely to its subsystems (expansion of the gullet). By contrast, protein transfer through the membranes of the paramecium is not a process that engages the unified behavior of the whole animal. (Burge, 2009, p. 263)

While Burge counts paramecium motility as involving action and therefore agency, he does not count tropisms or prokaryote motility as manifestations of agency, and so organisms such as bacteria and plants do not count as agents. About tropisms, he writes: “In many cases, tropisms are nothing more than oriented growth. I lay tropisms aside. They are mostly either non-active movement or at best borderline cases of active movement,” (Burge, 2009, p. 257). He does not count prokaryotes like E. coli as capable of agency because they are not able to directly control the direction in which they propel themselves. By changing the frequency with which they “tumble” instead of “run,” E. coli are able to significantly increase the likelihood of swimming towards food sources, but they do not actually steer themselves.

Moreno and Mossio criticize Burge’s account, writing that Burge “seems to rely more on intuition than on a rigorous conceptual base” (2015, p. 95). With the present framework, we can go further than this and understand the difference between paramecium and prokaryote motility in terms of the capacity for *targeting* (Section 4.6.3.3), which I defined as the minimal requirement for weak goal-directedness. The difference between central and peripheral capacities that Burge relies on more generally, however, does not seem explainable in these terms.

Above, in Section 5.4, I defined a distinction between the “central” and “peripheral” controllers that are part of the larger composite control system which is the agent itself. The difference between the central and peripheral controllers is that the input and output repertoire of the central system form what Grush and Springle referred to as an “inverse intersection” (forthcoming, p. 9), a common repertoire of contentful states that represent states of affairs. The peripheral systems maintain these repertoires using information sources internal and

external to the organism and translate output repertoire selections into fully articulated behaviors or motor commands. While this is a rigorous and “objective” (in the sense of being grounded in the system’s own perspective, rather than the perspective on an external observer) way of distinguishing between central and peripheral systems, it would of course not serve Burge’s purposes since the information processing capabilities of single-celled organisms certainly do not have the sophistication necessary for representation of states of affairs *as* states of affairs.

Burge’s notion of agency, unlike the present one, is not derived from presupposing a meaningful distinction between *choices* and *selections*, and seems to be more concerned with capturing casual usages (of biologists among others) of the word ‘action’. As with the distinction between paramecium and prokaryote motility, it is likely that in particular contexts, attribution of activities as ‘actions’ will turn on the presence or absence of some control system feature or other (such as those described in the previous, present, and following chapters), but it is unlikely that usage of ‘action’ or ‘agent’ will be captured across the board by any particular set of control system features. This is why the main value of the present framework lies in its laying out of the various features that are likely to be important for understanding such usages, rather than in legislating about the language itself.

Before moving on from the discussion of Burge, one further comment he makes deserves attention. While Burge’s account of agency is much thinner than the one I’ve offered here, his account is also very different from mine in that he also argues that agency does *not* necessarily involve control:

> Primitive whole-organism agency often involves whole-organism *control*, but does not require it. Ducking an approaching missile can be an action even if it is against one’s own attempt to inhibit the ducking. I assume that the ducking is not a peripheral reflex in the classical sense. It is guided by perception. Such ducking seems intuitively not to be under the individual’s control. The individual would naturally say that he or she could not help but duck despite trying not to. ... A more fundamental reason against taking control to be central to primitive agency is that with respect to the simplest organisms, the notion of control has little grip. Primitive whole-organism agency also does not require a capacity to *shape or*

> *guide whole-organism movement* past the point where the stimulus is registered. Various types of instinctive behavior are inflexible and chain-reflexive, but still count as action. The male grouse will copulate with a stuffed grouse, male grouse, or dead grouse, if it sees any of these as assuming the relevant female mating position. The male grouse's copulation activity is released by a single stimulus or single perception. The instinctive behavior does not derive from an inability to distinguish visually between the sexes, or between live and dead grouses. It is just that the instinctive behavior overrides these distinctions, once the key stimulus is received. … Intuitively the grouse's action and the ducking are guided by the individual's perception. But the action is not under the control or guidance of the individual in the sense that the individual need not endorse the behavior and could not monitor or adjust it, given the initial perceptual input. (Burge, 2009, pp. 264–265)

Here, it is apparent that Burge understands control in a different way than the account I offered in the previous chapter. The phrase "he or she could not help it" usually refers to a type of control that philosophers refer to as *intentional control*, which involves the ability to prevent behaviors that deviate from one's intentions. As I mentioned in Section 4.7.4, the nature of intentional control is a topic that has been highly debated by philosophers who are interested in defending causal theories of action from the problem of deviant causal chains.[93] As long as 'control' is being understood in the sense of *intentional* control, Burge is of course right that "primitive agency" (as he understands it) would not necessarily require control, since for him, primitive agents may not have intentions at all.

The last sentence of the quotation suggests a weaker understanding of control, however. Instincts can give rise to *ballistic* behaviors that are not guided by feedback once they are initiated (Klinger, 1977, pp. 84–85; Lea, 1984, p. 22). Such behaviors are generally aimed at a goal at least when or before they are initiated, and still count as what I referred to as "weak goal-directedness" in Section 4.6.3.3. There, I also distinguished between guided and unguided targeting; unguided targeting corresponds to aimed ballistic behaviors. Both types of behaviors ultimately involve control, of course, in the sense I defined in Chapter 4; the difference is in the

[93] I do not necessarily consider the account developed in this chapter as a version of the "causal theory of action." As Frankfurt wrote, "Explaining purposive behavior in terms of causal mechanisms is not tantamount to propounding a causal theory of action. ... the pertinent activity of these mechanisms is not prior to but concurrent with the movements they guide" (1978, p. 160).

type of control involved. As long as 'control' is being understood in the sense of *guided targeting*, again, Burge is right that "primitive agency" (again, as he understands it) would not require control, the copulation activity of the grouse being a case in point.

5.6.3 Steward

Helen Steward (2009, 2012) offers an account of minimal agency designed to accommodate the attribution of agency broadly to animals as well as humans, on which:

> (i) an agent can move the whole, or at least some parts, of something we are inclined to think of as *its* body;
> (ii) an agent is a centre of some form of subjectivity;
> (iii) an agent is something to which at least some rudimentary types of intentional state (e.g., trying, wanting, perceiving) may be properly attributed;
> (iv) an agent is a settler of matters concerning certain of the movements of its own body i.e., the actions by means of which those movements are effected are considered to be non-necessitated events, attributed always first and foremost to the agent, and only secondarily to environmental impacts or triggers of any sort. (2009, p. 226)

About condition (i) she writes:

> [W]e immediately impose, for example, on our conception of an animate entity, at least at first, [a] dualistic scheme of animal and body, according to which we suppose the animal to be a possessor of its body, in much the way that we are possessors of our own. There is the cockerel, and there is the cockerel's body which the cockerel can make move, just as we can make our bodies move. (2009, p. 225)

The idea here seems to be that there is a conceptual distinction made between the agent *as controller*, and the agent (or its body) *as controlled*. This appears to me to correspond to the basic distinction between the controller and the controlled system (or "plant") in control theory. I therefore interpret condition (i) as a corollary of the analytically true fact that, as Rescher pointed out (Section 4.4), in any instance of control, there must always be something doing the controlling, and something being controlled. (i), then, follows from the fact that an agent is a type of control system.

On condition (ii), Steward does not provide much in the way of elaboration, except to characterize it as "the presupposition that an animal is a centre of some form of subjectivity, subjectivity which affects and mediates its interactions with its environment" (2009, p. 225). The present view can provide substantiation for this criterion, however, since it holds that agency involves having a perspective on the world as being carved into a number of kinds of states of affairs, as well as a perspective on what states of affairs currently obtain and which ones are under the agent's control. The agent's repertoire of types of states of affairs, beliefs, and preferences together constitute a categorization scheme and normative standard for how it should influence its own environment.

Condition (iii) is intended to be thinner than it sounds. First, Steward is clear that she does not take "intentional states" to necessarily be propositional attitudes. But further, she seems to mean by "intentional states" what I have referred to as strong goal-directedness (Section 4.6.3.4): the ability to combine representational models of the environment with representations of value (what she calls "desire-like states") that are then used to drive behaviors. This would not require the capacity for objective representation of states of affairs that I have described in this chapter, and therefore would not require the kind of distinction that I have emphasized between choice and selection. It would also not be sufficient to allow for the kinds of reasons explanations that I discussed above in Section 5.2.2 (except from the context of an external observer's *intentional stance*, but Steward, 2012, p. 107 rejects instrumentalism about agency).

On the other hand, Steward claims that she intends condition (iv) to capture the capacity for *choice*, writing that

> Our natural inclination is to think of an animal as a creature that can, within limits, direct its own activities and which has certain choices about the details of those activities. To invoke a terminology I find helpful in this context, it is natural to think of animals as the *settlers* of various matters which concern the movement through time and space of their own bodies, and I submit that it goes deeply against the grain to suppose that each exact detail of each movement

> orchestrated by an animal was settled at any point prior to a period broadly coeval with what we think of as the period of the animal's action. (2009, p. 226)

Recall from Section 4.3 that Rescher characterized control in general as the capacity to "affect decisively" an outcome, and even to "assure" a "desired result." There are, then, a couple of ways to interpret the intuition that animals are the "settlers" of various matters. Any autonomous controller has a perspective on when it should have one kind of influence on a control parameter rather than another; animals are no exception. I argued in the last chapter that this constitutes a rudimentary kind of normative standard, the kind that can be invoked whenever literal selection is said to occur. In Pattee's (1973b, p. 71) terminology, any selector or controller must incorporate "decision-making constraints."

The other way to interpret the idea of animals as the "settlers" of matters concerning them is to consider them to be capable of choosing in Daveney's sense. I have argued, however, that the representational requirements for this go beyond strong goal-directedness and require *preferences*, which in turn require the ability to represent states of affairs *as* states of affairs, decoupled not merely from particular behaviors (Sterelny's sense of "decoupled representations), but further decoupled from *control* as such (i.e., decoupled from direction of fit). In this way, the agent does not merely make a selection (or "pick" an outcome), the agent selects based on which *outcome* the agent itself considers to be the *best*. Depending on how condition (iv) is read (or intuited), then, it may constitute an argument in favor of the view of agency I have put forward.

Ultimately, Steward's views about agency are designed primarily to facilitate her larger (2012) metaphysical position on the freewill debate, in particular, her response to Van Inwagen's (1983) "consequence argument." She attempts to formulate a libertarian metaphysics that incorporates a non-perspectivalist realism about agency and non-deterministic agential causation. Since my view combines internal perspectivalism about both control (Chapter 4) and agency (Section 5.2.2), multi-perspectival realism about mechanistic

causation (Chapter 2), and the view that controllers and agents are types of mechanisms, it would likely be considered largely incompatible with her larger metaphysical goals.[94]

5.6.4 Moreno and Mossio

Moreno and Mossio's (2015; henceforth M&M) account of agency is the most fully developed of a larger class of accounts that define agency in terms of self-maintenance (e.g., Bickhard & Terveen, 1995; Christensen, 1999; Di Paolo, 2005; Lyon, 2006; Thompson, 2007; Campbell, 2009; Hooker, 2009; Witherington, 2011; Arnellos, Bruni, El-Hani, & Collier, 2012; Walsh, 2015; Jones, 2017). M&M, like many other authors who develop self-maintenance-based accounts, are primarily concerned with developing their notion of biological autonomy and arrive at their notion of agency within the context of this broader project.

M&M's account begins with the plausible claim that the concept of agency inherently involves a differentiation between a self (the agent) and its environment (2015, p. 91). Such a differentiation naturally falls out of their conception of biological autonomy, given that self-maintenance is essential to the latter (without a self with its own identity distinct from its surroundings, *self*-maintenance could not have meaning). M&M claim that a second feature that follows from the very concept of agency is that agents have a "capacity to generate causal effects: agents are the source of interactions that are not determined by either the events of the immediate or distant past, or by physical laws of nature" (2015, p. 92). They explain that "agents are the sources of causal effects because these effects are generated by the constraints that belong to their organization" (ibid.). M&M are working with a notion of

[94] In a future paper I will combine the present framework with an elaboration of MacKay's (1962) control-theoretic approach to the freewill debate. The basic idea is that microphysical determinism and agent causation are distinct but equally valid perspectives that can be taken on the non-perspectival reality of modal patterns that ground both perspectives.

‘constraint’ that has descended from the work of Howard Pattee, similar to that developed in Chapter 2 of this dissertation.

Third, M&M write that agency has “teleological and normative dimensions” in that “actions are supposed to have goals and comply with norms” (2015, p. 93). The details of their account are then derived from these three initial claims, which may be restated as follows:

(1) Agency inherently involves a differentiation between a self (the agent) and its environment.

(2) Agents initiate and determine at least some of their interactions with their environments.

(3) The interactions that agents initiate are imbued by the agent with normativity.

From only these basic principles, M&M derive their definition of agents as those systems that maintain their own organization partly by initiating and determining their interactions with their environments “in a teleological and normative way” (2015, p. 93). How do M&M arrive at this definition?

In an earlier chapter of their (2015), and as discussed in Section 3.3.3 of this dissertation, M&M defend a self-maintenance-based account of proper function normativity: traits are functioning properly when they are actually contributing to the self-maintenance of the larger organized system. M&M then combine this with statement (3) to make an argumentative move that one commonly finds among the larger family of self-maintenance-based approaches to agency listed above—an appropriate name for it might be the “Standard Argument for Agency as Grounded in Self-Maintenance”:

> By contributing to the maintenance of the closed organisation to which they belong, agential functions contribute to maintaining the conditions of their own existence; hence, the maintenance of the whole organisation can be taken as the naturalised goal of agential functions, and its conditions of existence are the norms of their activity. (2015, p. 93)

I noted above that the self-maintenance-based approach provides a natural way to accommodate statement (1). To see how it is intended to accommodate statement (2), recall

M&M's claim that "agents are the sources of causal effects because these effects are generated by the constraints that belong to their organization" (2015, p. 92). M&M argue that their account of self-maintenance as involving a certain kind of organization called "constraint closure" results in the emergence of causal powers (2015, p. 51). On their view, it is then these very causal powers that allow agents to initiate and determine their interactions with the environment.

Only those systems that realize constraint closure have the kind of causal independence that counts as agency. Based on their definition, M&M argue that chemotactic bacteria and higher organisms, but not viruses, count as agents (2015, pp. 96–98). This is because whereas for bacteria like E. coli, the organization necessary to initiate complex interactions with the environment such as chemotaxis lies within their own self-maintained boundaries, viruses

> do not possess a constitutive organisation that is complex enough to perform agential capacities: they exhibit such capacities only insofar as they are integrated into much more complex systems (typically: cells) that are organised, in the specific sense that they realise a closure of constraints. (Moreno & Mossio, 2015, p. 96)

There are 5 main ways that the framework developed in this dissertation differs from that of M&M that are relevant for evaluating M&M's account.

First, unlike many authors (including M&M) that adopt self-maintenance-based accounts of agency, I do not consider the normativity of agency to be derived from the normativity of proper function. I argued in this chapter that agents are control systems, and their normativity derives from having a perspective on what states of affairs are under their control at what times. In Chapter 3, I considered and rejected the self-maintenance-based account of proper function and instead argued that the normativity of proper function is based on a system having a perspective on how its components (the "beta creatures") are supposed to perform. In Section 4.5, I argued that control systems do not necessarily have this type of perspective; proper function normativity is orthogonal to control normativity. Having said that,

some forms of control do incorporate proper function normativity, for example in reinforcement learning, certain forms of strong goal-directedness, and the use of Kalman filters with forward models (on the latter, see Grush, 2009). In such forms of control, some component of the input/output mapping will have a certain role, and the control system then modulates the input/output mapping on the basis of whether that component is detected to have played that role correctly or not. This type of normativity is not essential to how I have defined agency in this chapter, however (though as a contingent matter, it may be present in all existing control systems that count as agents on my definition).

Second, I do not consider the normativity of control to derive from self-maintenance. A system need not be self-maintaining to be an autonomous controller in the sense defined in Chapter 4. What makes a controller an autonomous controller is the fact that it operates with its own categorization scheme for sensory and control parameters, and its own mapping from sensory to control parameters. The system's influencing of the control parameters need not result in self-maintenance. In fact, I argued in Section 3.3.3 that the attempt to ground a normative standard in self-maintenance fails due to generality problems. The account of control and agential normativity I am adopting here is instead internal perspectivalist, and so is not vulnerable to generality problems in the way that self-maintenance-based views are.

Third, on the framework I have developed in this dissertation, the emergence of causal powers does not require constraint closure. In Chapter 2, I provided an account of the emergence of directional causal powers based on a certain kind of second-order structuring of constraints, but this structuring did not require a level of complexity remotely approaching autonomous self-maintenance or constraint closure. Similarly, as a fourth point, possession of active causal powers does not require agency on my view. I instead defined systems capable of realizing proper function, control, and agency in terms of mechanisms and their directional causal powers. Further, not all mechanisms are intrinsically normative: as I argued in Chapter

4, only those mechanisms that are *observer-worker systems* realize any kind of normative standard by themselves.

Fifth, I don't think that a system has to have a particular perspective on where its own boundaries begin and end in order to be an agent. Agents will often be capable of sensing both internal and external states (called interoception and exteroception), and will respond differently to internal states than to external states. Further, agents often respond differently to sensory features that are produced by a part its own body (for example, a dog whose vision is partly occluded by its snout does not behave as if an object is close to its face or getting in its way). But this does not require the agent to identify such internal states and body parts producing certain sensory effects as states *of the same being*, or to in any way identify that being as *what does the controlling*. In fact, most organisms likely have no perspective at all about their own "self" as a unified object distinct from its environment; they would not, for example, understand that the image they see in the mirror is identical to what is doing the seeing (or, the same dog that seems aware of its own snout may start chasing its tail). I do not, therefore, think that statement (1) lends support to a self-maintenance-based view, or to any view of agency as requiring the agent itself to have a perspective on where its own boundaries are. As I argued in Chapter 3, however, living biological systems do, in general, have such perspectives that are instead subagential and grounded in proper function normativity.[95]

[95] Grush (2000) discusses the role of allocentric representation in model-based control, which may require registering the bodily "self" as an individual distinct from other individuals in one's model. As Grush (2009, p. 311) points out, however, this type of representation does not necessarily constitute a perspective on where the boundaries of the agent itself lie. Such a perspective would arguably require the agent to attribute its own controlling activities *to that bodily self*, which would in turn involve much more sophisticated representational capacities.

5.6.5 Dretske

According to Dretske (1999), an agent is a system that is capable of behavior caused by the system's own internal representational states, such that the *meaning* of the internal states explains why they cause the behavior. Dretske argues that a control system like a thermostat is capable of behavior that is caused by an internal representational state. An internal component—in Dretske's example, a bimetallic strip—represents whether or not the temperature rises above a certain threshold by whether or not it bends enough to complete an electric circuit, and the furnace turns on as a result. But Dretske argues that the thermostat is not an agent because

> The internal element whose degree of curvature means something is causally active, yes, but the fact that its curvature means something about temperature is not a fact about the cause that explains the effect. Meaning is there, but it is not doing anything. (1999, p. 23)

In other words, it is the mere fact that the strip has bent enough to complete the circuit that causes the furnace to turn on, and this would occur even if the strip did not bear a representational relation to the temperature. For example, one could also bend the strip with a pair of pliers. Dretske argues, therefore, that the representational property of the strip does not explain why it turns the furnace on.

Dretske acknowledges, however, that the representational capacity of the bimetallic strip "indirectly" explains why it was incorporated into the thermostat in the first place:

> Meaning explains – through us – why the cause is having this effect. It thus explains, indirectly, why the effect is occurring. We put the metallic strip there, we gave it that job to do, because of what its curvature means about the quantity (temperature) we want controlled. Given our purposes (to control room temperature), had the metallic strip's curvature not been a reliable indicator of temperature, it would not have been made into an electrical switch for the furnace. This internal element's curvature is causing what it does – and, thus, the thermostat is behaving the way it is – because this curvature means what it does. In this indirect way, then, meaning becomes explanatorily relevant. It achieves its relevance *through us*, the designers and makers of instruments. (1999, p. 24)

The key to focus on in this quotation is the claim that “we gave it that job to do.” In other words, it is only in the perspective of external observers that the component has its “job” of serving as a temperature-sensitive switch for the furnace. And it is only by virtue of its possession of this “job” that it becomes true to say that the meaning of the curvature explains the furnace’s turning on.

By contrast, animals are capable of acquiring representational states in such a way that their behavior becomes explainable directly in terms of the meaning of those states. Dretske considers a case of devaluation similar to the one in Dickinson and Balleine’s experiment discussed earlier:

> A foraging bird tries to eat a Monarch butterfly. This butterfly has been reared on a toxic form of milkweed. Such butterflies are poisonous and cause birds to vomit. After one nasty encounter, the bird avoids butterflies that look like the one that made it sick. A day later our bird sees a tasty Viceroy, a butterfly with an appearance remarkably like that of the noxious Monarch. The Viceroy, though, is not poisonous. It has developed this coloration as a defense from predatory birds. It mimics the appearance of the Monarch so that birds will "think" that it, too, tastes nasty and avoid it. Our bird sees the Viceroy and flies away. (Dretske, 1999, p. 27)

In this case, Dretske argues, the bird’s representation of the butterfly has acquired a new meaning, a nonconceptual representation of something like “nasty tasting bug.” Having acquired this new meaning, the representation then causes the bird to avoid the butterfly. Dretske argues that unlike with the bimetallic strip, this representation is a case where the meaning directly explains the behavior; this explanatory role is not mediated by the perspective-taking of an external agent. Like Anscombe (1963, pp. 7–9), Dretske considers behavior explainable in terms of the *agent’s own* reasons to be the essential distinguishing characteristic of agency as opposed to other systems capable of behavior (like human-built thermostats). After learning from the previous experience of eating the Monarch, the bird now has its own reason for avoiding similar butterflies, and its behavior is “governed by” this reason (Dretske, 1999, p. 30).

Why, exactly, is the bird acting on its own reason, but not the thermostat? There is presumably something in the bird's brain that is analogous to the bending of the bimetallic strip. One might argue that though the strip historically acquired its role in a different way from the bird's representational state, it ultimately is playing a similar type of causal role in the moment that it is actually causing the behavior. Dretske also considers the possibility of an organism having representational states with a meaning acquired not through learning, but through natural selection. Dretske's example here is the Scarlet Gilia, which changes its color from red to white at the same time each year, around mid-July. In this case, the plant possesses something like the bimetallic strip, a chemical clock that acts as a switch at the appropriate time of year. Dretske argues that such a feature may be the result of natural selection, so that the chemical clock has acquired the role it plays *because* of its ability to affect the flower's behavior based on what it represents. But unlike the thermostat, the explanatory role of the clock's representational contents is not mediated through an external observer.[96]

Dretske argues that even if this natural selection story is true, and even if the plant's changing colors can be construed as "behavior" caused by the chemical clock, the flower, unlike the animal, is not an agent, and the flower's changing color, unlike the bird's avoiding the Viceroy, is not an action. Dretske explains as follows:

> Unlike the thermostat and the plant, though, the meaning of the bird's internal representation (of the butterfly) is directly relevant to its behavior. Like the thermostat and the plant, this internal representation (call it R) has both a meaning and a causal role, but, unlike the instrument and the plant, its meaning explains its causal role. R is causing avoidance behavior, it was given that job to do, because it means that a butterfly of type M is present, the sort of object the bird, after its unpleasant experience, wants to avoid. So the causal story looks like this: an R which means M causes avoidance because it means M. A meaningful state is not only causing behavior (this was also true in the thermostat and the plant), its meaning explains why it is causing it. Meaning is thus explanatorily relevant to why the bird is behaving as it is. (1999, p. 29)

[96] Arguably, "natural selection" should not be construed as something that has a determinate *perspective*; see Section 3.3.1.

Though Dretske does not emphasize it, the key part to focus on is the phrase “it was given that job to do, because… .” Again, what allows representational meaning to have a causal explanatory role is not merely *that* is has been given a job to do so, but *who* or *what* gave it that job. Ultimately, the point here is that like Moreno and most of the self-maintenance theorists about agency, Dretske is grounding the normativity of agency (i.e., the fact that agency involves behavior that is explainable in terms of the behaving system’s own reasons) in the normativity of proper functions. A representational state can have a meaning that serves as an agent’s *own reason for acting* only if this meaning is grounded in a “job”-possession perspective (i.e., a proper function perspective) that is *internal to the agent*.

Thus, Dretske’s account of agency is centered around what amounts to an internal perspectivalist account of proper functions—not internal perspectivalism about proper functions across the board, but only about the kinds of proper functions that allow representational states to play the role of agential reasons. This explains why learning history is so important to Dretske’s account. The connection between reinforcement learning and internal perspectivalism about proper function can been seen by the fact that reinforcement learning is often described in terms of an “actor-critic” model, as Sutton and Barto explain:

> Actor–critic methods are TD [temporal-difference] methods that have a separate memory structure to explicitly represent the policy independent of the value function. The policy structure is known as the actor, because it is used to select actions, and the estimated value function is known as the critic, because it criticizes the actions made by the actor. Learning is always on-policy: the critic must learn about and critique whatever policy is currently being followed by the actor. The critique takes the form of a TD error. This scalar signal is the sole output of the critic and drives all learning in both actor and critic… (1998, p. 151)

From this description it can easily be seen that reinforcement learning (at least on the actor-critic model) involves the generation and enactment of proper function norms in just the way described in Section 3.5.3. The “actor” corresponds to what I referred to there as a “beta creature,” and the “critic” corresponds to an “alpha creature.” The actor-critic model is a little more sophisticated than the picture laid out in Chapter 3, however, because it provides a way

for the performance norms to shift over time, since the behavioral policy is independent of the "value function" (which itself might involve valuations of behavior outcomes).

While many control systems will of course incorporate something like the actor-critic model as part of how their input/output mapping works, it is important to emphasize again that the basic normativity of control is distinct from proper function normativity (Section 4.5). Proper function normativity involves a perspective on what it is the *job* of some part of a larger system to do. Control normativity involves a perspective on how something should be influenced over time. While it may be reasonable to conclude that "alpha creatures" that do the censoring that partly realizes norms of proper function acts as a *controller* for a given beta creature, the resulting control norms are not identical to the proper function norms realized by the larger system that potentially includes many alpha creatures and many methods of censoring. Censoring can be an instance of control but the two should not be conflated.

Further, I have argued that agency requires preferences, which involve a perspective on which some things can be under my control at some times but not at others. What it means for an agent to act on its own reasons is that such preferences are used in making choices, which then result in the behavior. This is not necessarily a proper function perspective: there doesn't need to be anything like an "alpha creature" or a "critic" that, itself, has a perspective on how the belief and preference states are themselves supposed to function, or on what their "job" is. All that is required is that the system is organized such that they in fact play that role. While looking for something internal to animals that can realize a form of normativity not present in plants and machines, I believe that Dretske has fixed on a real form of normativity realized by an internal perspective, but not the form of normativity that is constitutive of animal agency.[97]

[97] It is arguably also not a form of normativity exclusive to animal agents, since robotic control systems can be made to incorporate actor-critic-based reinforcement learning (e.g., Muse & Wermter, 2009).

5.7 Conclusion

In this chapter I have offered an account of agency that is not meant to be the one true definition of agency, but instead intended to serve as a fairly demanding reference definition that other notions of agency can be defined in relation against. This account is built around the central idea that the capacity, not merely to behave in a functional or adaptive way, and not merely to make selections, but instead to make *choices*, is the feature that workers across various disciplines are most often interested when talking about agency as a distinct type of organization. I have introduced a set of terminology as follows:

Agent. An agent is a control system whose input/output mapping incorporates preferences about objective states of affairs in order to make choices.

Choosing (5.2.1). Choosing is making a selection of an objective state of affairs against other objective states of affairs, on the basis that the selected state of affairs is considered best from among the options. This requires representing state of affairs *as* objective states of affairs.

Representing something as objective (5.2.2). Objective representation, i.e., representing states of affairs as states of affairs, involves having a perspective on something as being potentially under my control sometimes and not under my control at other times.

Considering a state of affairs to be the best from among the options (5.2.3). What determines whether a state of affairs is being chosen based on its being considered to be the best is whether or not it is chosen because it is *preferred* over other states of affairs.

Preference (5.2.3). A preference is a certain kind of modification to a control system's input/output mapping. It is like an implicit rule (in the Sellarsian sense of "rule-governed behavior"; see Section 3.5.2) that is operative in a control system

that says that when presented with certain alternatives, one of them should be selected instead of the others (where 'should' is being used in the same razor-thin sense in which the alpha creatures from Chapter 3 operate with an implicit rule that beta creatures "should" be censured under certain conditions). But it is an *effective* rule: while it has the preference, the preference is an operative part of the input/output mapping of the control system. Further, the alternatives to be selected between are states of affairs. Possession of this implicit rule requires that the control system must be able to *recognize* the presence of an alternative of states of affairs that can be selected between.

Recognizing the presence of an alternative of states of affairs (5.2.4). This requires recognizing that several competing states of affairs as the outcomes of competing behaviors that could potentially be chosen. This in itself requires a set of capabilities that only certain control systems and organisms possess. Namely, the organism/control system must be capable of detecting the presence of some state of affairs S and of anticipating the fact that a certain behavior B1 will causally produce some state of affairs R1, and an alternate behavior (or lack of behavior) B2 will result in some state of affairs R2 (S might be identical to or represented as identical to R1 or R2, but R1 must be represented as distinct from R2 for one to be preferred, of course). It is then R1 and R2 that are directly compared, not B1 and B2. Either B1 or B2 is selected based on whether R1 or R2 is preferred.

I have tried to show the place that a capacity for making choices occupies in a larger framework of capacities that includes other varieties of control, such as negative feedback, model-based control, weak and strong goal-directedness, and the possession of sub-agential motivational systems (a.k.a. "drive systems" or "motives"). I have also distinguished agential normativity from other types, such as basic selector normativity, proper function normativity, basic control

normativity, and the “actor-critic” normativity often involved in reinforcement learning. I have also defended the idea that agency, being a form of autonomous control, is grounded in the perspective of the system itself; an agent can have a perspective on states of affairs as being sometimes under its control and sometimes not.

I then showed how “folk psychological” states such as preferences, beliefs, and desires emerge from such perspective-taking activity. The use of preferences in conjunction with such a sophisticated representational system as described in Section 5.2.4 is what makes “agential states” like beliefs and desires necessary. As explained in Section 5.3, beliefs and desires prevent the agent from having to hold all of its preferences “in memory” all the time, which would require brains to be much larger than they are. The function of beliefs and desires is to enable the generation of preferences that can support the above representational capabilities on the fly. In the next chapter, I build on this account by dividing intentions, beliefs, and desires into two classes: inclinational and committal, and show how committal agency builds on the foundations laid in this chapter.

Chapter 6
Inclinational and Committal Agency

> That was the moment I made the decision. It was like I had stepped through a door and locked it behind me.
>
> —Capt. Benjamin Sisko, "In the Pale Moonlight," *Star Trek DS9*

6.1 Introduction

Whereas non-human animals are generally limited to doing whatever they feel most *inclined* to do at a given moment, humans are distinguished by their ability to override these inclinations and instead act on their *commitments* (e.g., commitments to goals or to policies). An example of this might be a woman who acts on her psychologically internal commitment (expressed in ordinary speech as what she has "decided to do") to going to work rather than her inclination (expressed in ordinary speech as what she "felt like doing") to go to the beach. For this reason, Klein, Molloy, and Cooper write that "commitment is a fundamental concept for understanding human behavior" (2009, p. 3). Commitment—like the topics of the previous chapters: constraint, function, control, and agency—is a highly multidisciplinary concept. The notion of commitment has been thought to be indispensable for addressing philosophical topics such as

- Free will (Holton, 2009)
- The distinctiveness of human action (Bratman, 2000)
- The self (Shoemaker, 2003)
- Personhood (Bratman, 2000)
- Akrasia (Shoemaker, 2003; Dodd, 2009)
- Self-control (Henden, 2008)

- Rational choice theory (McClennan, 1990; Peter & Schmid, 2007)
- The nature of concepts (Brandom, 2000)
- Altruistic motivation (Sen, 2005)
- Moral responsibility (Wolf, 1990)
- Scientific knowledge (Polányi, 1958, Chapter 10)

Further, commitment is a well-established construct in experimental psychology (Klein, Cooper, & Monahan, 2013; Allen, 2016; Klein & Park, 2016), and has been key to investigations into

- The self (Lydon, 1996)
- Well-being (Brunstein, 1993; Cantor & Sanderson, 1999)
- Effectiveness of psychotherapy (Hayes, Strosahl, & Wilson, 2012)
- Cognitive dissonance (Brehm & Cohen, 1962; Harmon-Jones & Harmon-Jones, 2008)
- Work motivation and organizational behavior (Locke & Latham, 1990; Klein, Becker, & Meyer, 2009; Meyer, 2016)
- Decision making (Janis & Mann, 1977; Montgomery, 1998; Baron, 2007)
- Emotional reactions (Klinger, 1987; Klinger & Cox, 2004)
- Representation of expected hedonic reaction to stimuli (Sharot, 2012)
- Perception (Lepora & Pezzulo, 2015)
- Attention and cognitive priming (Gollwitzer, 1999; Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trötschel, 2001; Huang & Bargh, 2014)
- Thematic content of thoughts and dreams (Klinger, 2013)
- Human sociality and joint action (Michael, Sebanz, & Knoblich, 2016).
- Persuasion and social influence (Cialdini, 2009, Chapter 3)
- Learning and educational outcomes (Morisano, 2013)
- The psychology of religious beliefs, experiences, and practices (Hinde, 1999)

Commitment is also an important concept in artificial intelligence and robotics, where it has been employed in research into

- Planning (Wooldridge, 2009)
- Communication (Chopra & Singh, 2013)
- Reasoning (Yolum & Singh, 2002)
- Cooperation and coordination between multiple agents (Jennings, 1993; Tweedale et al., 2007)
- Human-robot interactions (Curioni, Knoblich, & Sebanz, 2016)

Commitment is clearly of high importance for understanding many facets of agency across various disciplines.[98] But the same could be said about other concepts that are connected to agency, such as emotion, cognition, and consciousness. Why, then, devote an entire chapter to commitment, as opposed to these other topics? The reason is that commitment is the key architectural feature that brings with it a crucial new kind of agential perspective: the agent's *own agency* becomes one of the things that the agent can view as either under its control or not under its control in various ways at various times. The kind of perspective introduced in Chapter 5 then becomes applied to the agent itself in a recursive manner, as will become clear as the chapter proceeds.

In what follows I will refer to agents that are capable of acting on inclinations but not commitments as *inclinational agents*, and to agents that are capable of forming and acting on commitments as *committal agents*. Both inclinational agents and committal agents have the capacity for *inclinational agency*, but only committal agents have the capacity for *committal agency*. The discussion of agency in the previous chapter did not distinguish between inclinational and committal agency. The reference definition of agency I provided was intended

[98] Neuroscientists generally don't talk about 'commitment' per se, but they often investigate related phenomena under the rubric of "cognitive control" or "executive functions" (see Section 6.5).

to be neutral with respect to this distinction; inclinational and committal agents are both kinds of agents, but the account in Chapter 5 will be more directly applicable to inclinational agents, which generally have a simpler architecture. In this chapter, I will explain why commitment is necessary for understanding agency in many contexts and provide an account of what sets committal agency apart from inclinational agency.

In Section 6.2, I explain why an account of agents like humans requires a notion of commitment. Here, I rely primarily on Bratman's (1987) account. Bratman's discussion of commitment is mostly limited to how it figures in intentions; in Section 6.3, I explain why agential states in general, such as beliefs, desires, intentions, and preferences, actually divide into two kinds: committal and inclinational. In Section 6.4, I turn to the question of how committal states are related to inclinational states and critically examine how some other philosophers answer this question. In Section 6.5, I offer some gestures towards a novel mechanistic account of how they might be related, drawing from Grush's emulator theory. Section 6.6 concludes by considering the sense in which the present account can be classified as internal perspectivalist.

6.2 The Importance of Commitment for Future-Directed Intentions

6.2.1 Why Future-Directed Intentions Are Not Reducible to Beliefs and Desires

In the previous chapter, I described agency as a kind of control system. Like any control system, there are input or sensory parameters, and there are control parameters. There is also a functional mapping between the two. With agents, this functional mapping is very sophisticated, and involves preferences. To act on preferences, the agent must have beliefs, and must be able to form intentions based on its beliefs and preferences. Desires can cause preferences, and vice versa. My preference for IPAs over lagers may result in a desire for an

IPA when I am presented with a choice between the two. My desire for an IPA may then lead me to prefer drinking out of a tulip class instead of a tall and narrow one. As discussed in Section 5.3.3, desires also play a coordinating role between beliefs and intentions. As relevant beliefs are updated, the desire causes intentions to be updated in real time. What if there are multiple desires that would result in a conflict if I tried to act on them simultaneously? Other things being equal, the stronger desire will generally "win out." The agent will act on the strongest desire, and not on other conflicting desires. Desires are often generated by motivational systems (Section 4.6.3.4) combined with associations formed through reward-based learning.

This is a simplified picture of agency that is often useful for characterizing animal behavior. Something like this picture of agency as dominated by beliefs and desires has often presupposed by philosophers of action. However, Michael Bratman has argued that this picture of agency is highly problematic when applied to humans, because it cannot properly accommodate the existence of *future-directed* intentions, something that was not addressed in the previous chapter.[99]

It is not hard to see that we appeal to future-directed intentions when explaining the behavior of agents. On Monday I informed you of my intention to go to Los Angeles on Thursday. A week later, after hearing that I did indeed go to Los Angeles on Thursday, you will feel satisfied that you can explain why I did so. If you hear that I went to Los Angeles on Wednesday instead, you will feel that an explanation is called for. Why would an intention formed on Monday explain behavior on Thursday but not the same behavior on Wednesday?

Philosophers have made a number of attempts to analyze future-directed intentions in terms of beliefs, desires, and the resulting present-directed intentions; Bratman includes Goldman, Anscombe, and Davidson (in earlier writings) on this list. Perhaps we should say that

[99] For a review of literature on the capacities and limits of "future-thinking" in non-human animals, see Redshaw and Bulley (2018).

my desire on Monday to go to Los Angeles was sustained for the next four days, and resulted in a present-directed intention to go on Thursday. But why did this desire only result in behavior on Thursday? Why didn't it cause me to go to L.A. on Monday? In answer to this question, we might suppose that the desire that was formed on Monday was for a temporally-indexed state of affairs: the situation of my being in L.A. on Thursday. This is tantamount to my having said to you on Monday, "What I want most is to go to L.A. on Thursday."

But can a desire that I had on Monday really cause me to go to L.A. on Thursday? Suppose that on Thursday I said "I don't feel any desire to go to L.A. today. But on Monday I did feel such a desire. So I'm going to L.A." This would not make any sense. Bratman argues, based on considerations like this, that future-directed intentions are a very different kind of psychological state than ordinary beliefs and desires. They are different because, unlike ordinary beliefs and desires of the kind that have been discussed so far, future-directed intentions involve *commitment*.[100]

In ordinary discussion contexts, one might say "On Monday I made a *choice* to go to L.A. today, and that's why I'm going, even though I don't feel the desire to go that I felt on Monday." There is a way in which this might seem to make sense, and a way in which it might not. Of course, once Thursday came around, one could choose at that point *not* to go to L.A., regardless of the choice that was made on Monday. In this sense, only the choice made on Thursday matters. On the other hand, sometimes when making a choice we simultaneously *commit to* that choice, which indicates the voluntary adoption of some kind of self-constraint. What is the nature of this self-constraint?

Before diving into this question, it is important to be clear about a couple of distinctions. First, Bratman differentiates between *inter*personal and *intra*personal commitment. Sometimes these are referred to as social versus psychological commitments, or public versus private

[100] Cohen (1992, p. 47) similarly distinguishes between beliefs and desires, on the one hand, which he claims do not involve commitment, and "having an intention" on the other, which does.

commitments. An interpersonal commitment is a commitment that one person makes to another, as in a promise. The second person may then hold the first person accountable for carrying out their promise. But this is not the same thing as an *intra*personal commitment. Often, one will make both kinds of commitment at the same time: one will internally constrain oneself to keeping a promise made to someone else. But an intrapersonal commitment does not require that one share one's commitment with anyone else.

Another important distinction is between external versus internal constraints on behavior. One may make a promise to others about their own behavior in order to add an external incentive that will reinforce their own behavior. Bratman provides another example, of "making a side bet with someone else that one will not gamble later on" (1987, p. 12). Such external constraints (called "precommitments" by Elster, 1979) do not have to involve other people: to prevent myself giving in to the temptation of eating too much junk food later, I might decide not to buy any while at the store. In discussing the kind of commitment involved in future-directed intention, Bratman is concerned with commitment that is both intrapersonal and internal in the sense that such constraint is not mediated by changes in one's external environment.[101] Similarly, McClennan argues that an account of rational agency must include a notion of commitment understood as the capacity for "endogenous preference changes" (1990, p. 215).

6.2.2 The Nature of Commitment-Based Self-Constraint

On Bratman's view, future-directed intention involves intrapersonal, internal self-constraint in two important senses: he refers to these as the "volitional dimension" and the

[101] Frankish refers to such intrapersonal and internal commitments as "effective commitments" (2004, p. 74n).

"reason-centered dimension" of commitment. By the "volitional dimension," Bratman means that intentions are, like desires, capable of initiating and energizing behavior.[102] In other words, intentions have motivational "oomph" (see Section 5.3.3). Bratman's way of putting this is to say that intentions and desires are both "pro-attitudes" (1987, p. 16). But here, Bratman argues that we must distinguish between two kinds of pro-attitudes. Whereas ordinary desires are "merely potential influencers" of action, intentions are "conduct-controlling" pro-attitudes. Bratman provides the following example:

> suppose I desire a milk shake for lunch, recognize that the occasion is here, and am guilty of no irrationality. Still, I might not drink a milk shake; for my desire for a milk shake still needs to be weighed against conflicting desires—say, my desire to lose weight. My desire for a milk shake potentially influences what I do at lunchtime. But in the normal course of events I still might not even try to drink a milk shake. In contrast, suppose that this morning I formed the intention to have a milk shake at lunch, lunchtime arrives, my intention remains, and nothing unexpected happens. In such a case I do not normally need yet again to tote up the pros and cons concerning milk-shake drinking. Rather, in the normal course of events I will simply proceed to execute (or, anyway, try to execute) my intention and order a milk shake. My intention will not merely influence my conduct, it will control it. (1987, p. 16)

The desire, then, will have the power to influence my behavior but may be in competition with other desires that may cause hesitation at the moment of action. A committed intention, on the other hand, will be much more robust in the face of competing desires. It will drive one's behavior without one needing to again consult one's desires.

Commitment also has a "reason-centered" dimension. If on Monday I form an intention to go to L.A. on Thursday, I will typically not continue to deliberate about whether to go to L.A. As Bratman writes, "my intention resists reconsideration: it has a characteristic *stability* or *inertia* … Lacking new considerations I will normally simply retain my intention up to the time of action" (1987, p. 16). Merely being confronted with competing, desirable options will not, by

[102] For an overview of empirical evidence of this aspect of intentional commitment, see Gollwitzer and Oettingen (2011).

itself, cause me to reconsider my intention; opening up my intention to reconsideration at all requires *persuasion*.

Further, committed intentions are normally taken for granted in one's reasoning processes. I may treat it as a foregone conclusion that I will go to L.A. on Thursday, so if a friend wants to meet with me in L.A. on that day, I feel free to form the intention to do so since I feel that the issue of whether I will be in L.A. on that day is resolved. Having a committed intention frees my cognitive resources up so that they can be allocated to deliberating about the *means* of executing my intention (e.g., what form of transportation to use, or what route to take). It will also constrain the formation of other intentions: I will not make other intentions that are inconsistent with going to L.A. on Thursday, like spending the day at the San Diego Zoo.[103]

Bratman argues that our ability to deliberate about and coordinate our activities, not only with our future selves but with other people, requires that we have the capacity for pro-attitudes with the above characteristics. Unlike other animals, humans are *planning agents* that need to be able to treat plans as *settled* without fear of our momentary desires interfering, and without having to continually hold all of the considerations that led to the formation of such plans in mind.

There is an additional sense, considered by Pepper (1958, pp. 79–84), in which a form of commitment is widespread in (non-human) animal behavior. Pepper discusses the fact that the strength of the energization of a given behavior will depend on not only on the strength of the underlying desire or drive, but also on the strength of the evidence that such behavior is based on:

> An animal typically gives less than his full commitment to an act when there is a conflicting alternative. In a dangerous situation an animal may move gingerly toward its goal. A dog with a bone moves past another dog with great circumspection. The dog is acting on the evidences of the total situation, and commits himself to the drive for depositing his bone only so far as he believes

[103] For overviews of empirical findings bearing out what Bratman refers to as the "reason-centered dimension" of commitment, see Harmon-Jones and Harmon-Jones (2008), Cialdini (2009, Chapter 3), and Klinger (2013).

> safe. There is an implicit judgment here with inertial commitment (which I think, we should agree in this instance, happens to be also intelligent). This instance shows clearly that animals can express doubt in their behavior. The doubt arises from conflicting impulses which inhibit the full discharge into the dominant drive. (1958, p. 83)

This is an important sense of "commitment," but it is different from the sense that Bratman is interested in. Commitment in Bratman's sense is not something that happens automatically as a result of momentary appraisal of evidentiary strength or potential for desire/drive satisfaction, but is instead the introduction of a pro-attitude that can override or modulate these momentary appraisals. An animal's automatic updating of momentary beliefs and desires[104] will of course act as a form of self-constraint, but not in the sense of overriding or controlling the animal's future momentary impressions and passions. In the rest of this chapter, I will be focused on commitment in Bratman's sense.[105]

6.3 Commitment as Not Limited to Intentions

6.3.1 Committal versus Inclinational Desires

Bratman generally describes intentions (especially future-directed intentions) as involving commitment, and desires as not involving commitment (in fact, Bratman refers to first-

[104] As de Sousa poignantly put it, "Bayesian decision theory is applicable to dumb animals" (1971, p. 57).
[105] For Polányi, these types of commitment are positions along a larger continuum:

> I have suggested before that in a generalized sense commitment may be acknowledged even at the vegetative level, since it is of the essence of a living organism that each part relies for its function, and for its very meaning as part of the organism, on the presence and proper functioning of a number of other parts. ... Commitment may then be graded by steps of increasing consciousness; namely, from *primordial*, vegetative commitment of a centre of being, function and growth, to *primitive* commitment of the active-perceptive centre, and hence further again, to *responsible* commitments of the consciously deliberating person. (1958, p. 363)

Examples of what Polányi calls "vegetative commitment" might include the process by which a bacterium "decides" to sporulate (Stephens, 1998) or the processes by which embryonic cells become irreversibly constrained to a certain developmental fate (Gilbert & Barresi, 2018, pp. 30–32).

order desires as “inclinations” at 2000, p. 38). But other authors have argued that in fact, desires break down into two kinds: ones that involve commitment, and ones that do not. According to Daveney (1961), there are at least three ways in which the word ‘want’ is used. He divides these into “intentional” wanting, “contemplative” wanting, and “inclinational” wanting. For example, suppose someone asks me why I am buying a train ticket, and I answer “because I want to go to L.A. on Thursday.” Daveney argues that in cases like this, ‘want’ is used synonymously with ‘intend’. In other cases, however, it is clear that ‘want’ and ‘intend’ can come apart. Daveney gives the following example: “Mr. Smith doesn’t intend to go to the meeting because his enemy Jones will be there, although he wants to go.” In this case, Daveney argues, the speaker is saying that Smith would go to the meeting under certain circumstances, in this case if Jones were not going. Daveney argues that ‘want’ here indicates a *conditional intention*, and he refers to these as cases of “contemplative wanting”:

> One may speak therefore of ‘want’ in the sense of an intention in cold storage, awaiting the occurrence of suitable circumstances. I call this sense of ‘want’ “wanting in contemplation”, simply because one contemplates what is wanted and is not engaged in its active pursuit. And the statement “I want …” I call an expression of *conditional* intention, as certain conditions have to be fulfilled before the agent is said to have the intention. This is quite different from “I intend to do so and so, if such and such occurs”, for here the intention is not conditional; it would be true to say the agent had the intention here and now, but the action he intends is conditional upon certain circumstances being fulfilled. By stretching the language a little it might be said that my conditional intention is what the agent *would* intend if certain conditions were fulfilled. But we don’t say “would intend”. We say “want”. (1961, p. 138)

Note that in Daveney’s example, there is not an existing intention that is overridden by a desire to avoid Jones. But Smith yet has an existing psychological state of being committed (in Bratman’s sense) to going if Jones isn’t there, and not if Jones is there.

Daveney distinguishes this from a third sense that he calls “inclinational wanting.” In this sense, it is possible to want to go to the dentist (in the contemplative sense) but to not want to go (in the inclinational sense). I may not feel any tooth pain, and I may (like many people) generally consider dentist appointments to be unpleasant. But I may also reason that it is in my

best interest to go to the dentist every so often. I therefore decide that I want to go to the dentist sometime. But this decision does not, by itself, yield an intention to go. I have not committed myself to going to any particular dentist (suppose I just moved to a new town) or to any particular timeframe for when I will go. This seems to correspond to Daveney's "comtemplative wanting," since I do not yet have an intention to go to the dentist, but I will when certain conditions are fulfilled. Perhaps at some point, after meeting some new people and asking them about local dentists, I will form such an intention. "Inclinational wanting," on the other hand, is a desire that is not the result of any such contemplation or commitment, but is instead the result of my momentary perceptual evidence, unconscious associations, feelings, urges, drives, etc. (i.e., my *inclinations*).

Something like Daveney's distinction between two kinds of 'wanting' seems to be operative in Watson (1975, p. 209) as well (though Watson reserves the term 'desire' for what Daveney calls "inclinational wanting"). Davis (1984) similarly distinguishes between "volitive" desires and "appetitive" desires. Whereas appetitive desires (corresponding roughly to "inclinational wanting") generally are accompanied with a physical urge or are associated with feelings of pleasure, volitive desires (corresponding to "contemplative wanting") are typically "based on *reasons*," by which Davis means that we typically have reasons consciously *in mind* for why we have the desire in question. Davis points out not only how volitive desires can come apart from appetitive desires, but also how they play a similar reason-centered constraining role to Bratman's intentions. Davis gives the example of someone who is asked "whether he wants to play tennis or golf at noon today," and then responds: "I desire to do both." This response makes sense if 'desire' is used in the appetitive sense, but not if 'desire' is used in the volitive sense. A volitive desire will typically constrain other reasoning and planning processes as well, even though it may not yet represent an intention (e.g., committing to a desire to play tennis instead of golf would not yet be tantamount to an intention to play tennis).

In what follows, I will adopt Daveney's term and refer to psychological states that do not involve commitment in Bratman's sense as "inclinational"; I will refer to states that do involve such commitment as "committal."[106] Instead of talking in terms of "ordinary desires" (Bratman's term), "inclinational wanting," or "appetitive desires," I will refer to *inclinational desires*. Instead of talking in terms of "contemplative wanting" or "volitive desires," I will refer to *committal desires*.

Thomas Nagel also drew a distinction between two kinds of desires, which he called "motivated" and "unmotivated" desires. As seen in the following passage, his "motivated" desires correspond to *committal* desires, and his "unmotivated" desires correspond to *inclinational* desires:

> many desires, like many beliefs, are arrived at by decision and after deliberation. They need not simply assail us, though there are certain desires that do, like the appetites and in certain cases the emotions. The same is true of beliefs, for often, as when we simply perceive something, we acquire a belief without arriving at it by decision. The desires which simply come to us are unmotivated though they can be explained. (1970, p. 29)

In fact, in Bratman's later work, he himself appears to make room for a kind of contemplative or volitive desire, writing that "to identify with a desire to A one needs actually to decide to treat that desire as reason-giving in one's practical reasoning and planning concerning some relevant circumstances" (1996, p. 197). Further, "One treats one's desire as reason-giving when one treats it as setting an end that can to some extent justify means and/or preliminary steps" (1996, p. 198). This would seem difficult to do consistently if one were not able to commit to it in a similar way to how one commits to an intention.

6.3.2 Committal versus Inclinational Beliefs

[106] Gollwitzer similarly used the word 'noncommittal' to refer to inclinational desires: "By forming goal intentions. people translate their noncommittal desires into binding goals. The consequence of having formed a goal intention is a sense of commitment that obligates the individual to realize the goal" (1999, p. 494).

As seen in the above quotation, Nagel also seems to allow for a distinction between committal beliefs and inclinational beliefs. de Sousa (1971) similarly distinguished between Bayesian belief, which does not involve commitment, and "belief proper," or "acceptance," which does (cf. van Fraassen, 1980, p. 88; Lehrer, 1989, p. 26). Instead of saying that belief proper involves commitment, however, de Sousa says that it involves *assent*. Similarly, Dennett (1978) argued that we should distinguish between *beliefs* and *opinions*. He uses 'belief' to denote the type of attitude, common among both humans and non-human animals, that is basically an impression or seeming that is currently guiding one's actions. Dennett follows de Sousa in appealing to the notion of *assent* to distinguish between the two kinds of attitude: For Dennett, what separates opinions and beliefs is that unlike beliefs, opinions involve assent. One makes up one's mind to have an opinion, but not to have a belief, because making up one's mind results in assent. Dennett points out that it is possible to have a belief p that guides one's actions, without knowing that one has that belief that p. But this is very different from having *decided* that p, believing p on the basis of *deliberation*. This latter sort of belief, Dennett argues, is on the contrary "something rather like commitment, rather like ownership" (1978, p. 303). Similarly, Baier (1979) distinguishes between what she calls "beliefs" on the one hand—which she says involve commitment—and what she calls mere "cognitive states," "registrations," or "Bayesian beliefs" on the other hand, which do not involve commitment.[107]

These authors are not alone in forming a conception of a belief-like attitude that involves commitment. Brandom (1994, p. 157; cf. Polányi, 1958, p. 28) in fact defines belief as a *doxastic commitment*, and he argues that it is because belief can involve commitment, that beliefs can play the kind of role that they do play in our deliberations and reasoning. In this

[107] Baier's notion of 'registration' is drawn from Bennett (1976, p. 56). Dennett (1978) is actually a response to an earlier draft of Baier (1979).

way, Brandom gives a similar role to commitment to that given by Bratman, although Bratman is concerned with commitments involved in intentions, not beliefs. For Dennett, whereas mere beliefs can be spontaneously generated through perception or introspection, one must be *persuaded* or *convinced* to assent to an opinion.

It is important to note at this point that Dennett's distinction between belief and opinion, which roughly maps onto de Sousa's distinction between belief and acceptance, is not the same distinction that Railton discusses between belief and acceptance. In "Normative Guidance," Railton writes:

> Although acceptance, like belief, can arise spontaneously, acceptance is much more amenable to volition and purpose, and hence more directly subject to decision. We do sometimes speak of *deciding whether to believe p*, but this is equivalent to *making up our mind whether p*. That is, the focus is on the question *whether p*—whether p is supported by the balance of evidence, intuitively plausible, etc.—while ignoring collateral effects attributable to the state of mind of *believing that p*. In contrast, *deciding whether to accept p* often is not equivalent to *making up our mind whether p*, and the decision typically focuses not only on *whether p*, but also on the costs and benefits of accepting or failing to accept p in the present context, many of which enjoy some independence from p's truth. (2006, p. 17)

So what we really have are three attitudes on the table: Dennett's 'belief', Dennett's 'opinion' (which corresponds to Railton's 'belief' and de Sousa's 'acceptance'), and Railton's 'acceptance' (which also appears to correspond to Stalnaker's (1984, pp. 79–81), Cohen's (1986, pp. 91–97), and Bratman's (1992) usage).[108] Dennett's notion of 'opinion', like Railton's notion of 'belief', involves making up one's mind whether p, so it is similarly non-amenable to volition[109] and purpose and represents a reason-centered commitment in Bratman's sense. We do not form opinions for specific purposes in specific contexts, just as Railton points out that we do not form beliefs for specific purposes in specific contexts. But we do sometimes *accept*

[108] For other references to similar usages of 'acceptance', see Frankish (2004, p. 81). Frankish (2004, p. 124) refers to 'acceptance' in de Sousa's sense as "doxastic acceptance", and to acceptance in Railton's sense as "non-doxastic acceptance." Frankish also refers to doxastic acceptances as "superbeliefs."
[109] Whether or not committal beliefs, or 'opinions' in Dennett's sense, are under voluntary control (a thesis sometimes called *doxastic voluntarism*) is a matter of debate; see Vitz (2008).

propositions for specific purposes in specific contexts (e.g., accepting a statement for sake of argument). In what follows, I will use 'committal belief' as synonymous with Dennett's 'opinion', and 'inclinational belief' as synonymous with Dennett's 'belief'[110] and Baier's mere 'registration'.[111]

6.3.3 Committal versus Inclinational Intentions

We saw that while Bratman described ordinary (i.e., inclinational) desires as potential influencers of behavior, he characterized intentions as "conduct-*controlling*" states that involve commitment. However, whereas Bratman is clearly concerned with committal intentions (as are many social psychologists such as Gollwitzer), other authors have argued for what we might call *inclinational* intentions. For example, McFarland writes,

> By intention I mean that there is an explicit (mental) goal-representation which is in some way instrumental in controlling the behaviour of the animal, or person. Thus if I have an intention to write the following sentence, then I have a mental representation of the goal-to-be-achieved, and this representation is instrumental in controlling my behaviour, in the sense that my progress in achieving the goal is compared with my representation of the goal. This is close to the everyday usage of this term, and different from the current usage of some philosophers. (1989, p. 124)

McFarland's usage corresponds more closely to how 'intention' was used in the previous chapter, and does not involve commitment in Bratman's sense (although it does involve commitment in Pepper's much thinner sense above). In McFarland's sense, desires in general that result in behavior will do so by means of a mediating intention. The distinction here corresponds to Pacherie's (2008) distinction between proximal intentions (which are present-directed and do not involve commitment) and distal intentions (which are future-directed and

[110] Though my usage of 'inclinational belief' will not carry with it Dennett's instrumentalism about psychological states.
[111] Gendler's (2008) 'alief' seems to be a subclassification of what I am calling 'inclinational belief'. Price (1969, pp. 205–206) refers to 'inclination' in describing the stages of belief-formation before one has made up one's mind, and uses the word 'commit' for when one's belief becomes settled.

involve commitment). Carruthers (2007) similarly draws from dual process theory (Wason & Evans, 1974; Evans & Over, 1996) to distinguish between “System 1 intentions” and “System 2 intentions.” Holton provides an example of what he takes to be a System 1 intention: “Seeing the stationary traffic ahead I form the intention to change lane, and start scanning the mirror for an opportunity to do so. If you ask me what I am doing I can tell you, but I have given no conscious thought to the matter” (2009, p. 53). Shpall distinguishes between “partial intentions,” which he characterizes as “mental states of *being inclined* to act,” on the one hand, and “full or outright intentions, which, as traditionally conceived, are mental states of *being settled* on acting” (2016, p. 817). Chang (2013) also distinguishes between non-committed intentions and commitments, arguing that the content of intentions often comes apart from the content of commitments that lead to such intentions. I will refer to intentions without commitment as “inclinational intentions,” and intentions with commitment (i.e., intentions in Bratman’s sense) as “committal intentions.”

6.3.4 Committal versus Inclinational Preferences and Choices

According to some authors, both preferences and choices can themselves be divided into two categories: some preferences involve commitment while some do not, and some choices result in commitment while some do not. First, preferences: Andreou (2007) distinguishes between “given preferences” and “chosen preferences.” Whereas given preferences are preferences that “an agent can just find herself with,” chosen preferences “can be conceived of as a system of ranking that the agent *commits to* in light of her preferences in the first sense and her choice situation” (2007, p. 119). Andreou uses this distinction to address a puzzle about how rational people can form intransitive preferences: rationality dictates that one form intentions that are consistent with one’s chosen preferences, not one’s given preferences. Similarly, Weirich argues that there is a sense in which preferences can arise as a

result of deliberation and making up one's mind, but denies that such preferences can be voluntarily chosen:

> The formation of the preferences is a response to deliberation and not an act of will, in contrast to a choice resting on the preferences. Forming preferences after deliberation resembles believing the conclusion of an argument that one sees is valid and that has premises one believes. The belief arises because of the cogency of the argument and not because of an act of will. Similarly, preferences arise because of the force of deliberations and not because of an act of will. (2013, p. 4043)

Experimental work also suggests a functional differentiation between committed and non-committed choices. For example, Polman and Russo (2012) demonstrate that when commitment to a preference is strengthened, without in any way altering the content of the preference or the subject's evaluation of the preferred option, this can lead to a difference in how the subject later processes information relevant to the preference.

Acts of choice, themselves, may be differentiated based on whether they result in the formation of a commitment. In some contexts, the word 'decision' is used to distinguish choices that result in a commitment (e.g., Frankfurt, 1988, p. 172; Shadlen & Kiani, 2013). Sen argued that economic theory must incorporate a distinction between choices that do not, and choices that do, result in commitment, writing that "commitment does involve, in a very real sense, counterpreferential choice, destroying the crucial assumption that a chosen alternative must be better than (or at least as good as) the others for the person choosing it" (1977, p. 328). Rather than following from one's preferences, Sen wrote, commitment is "closely connected with one's morals" (ibid., p. 329). One way to understand Sen's point is to assume that by "preferences" Sen is referring to one's *inclinational* preferences, whereas one's "morals" may be partly constituted by *committal* preferences.[112]

[112] Chang (2017) argues that reasons themselves break down into two kinds: "given reasons" and "will-based reasons." For Chang, will-based reasons, but not given reasons, derive from commitments. In the present terminology we might then refer to this as a distinction between *inclinational reasons* and *committal reasons*.

6.4 How are Inclination and Commitment Related?

Though different authors use different terminology, there is widespread consensus that commitment is crucial for agency at least in humans, and that commitments have motivating force, effects on cognition, and normative force (in terms of reasoning, planning, etc.).[113] Different kinds of psychological states (not just commitments) can be characterized in terms of the degree to which they have the dispositional nature and normative force of inclinations or of commitments. But there has been little consensus about how the relation between inclinations and commitments should be characterized. I first consider several prominent positions on how they are related, before turning to my own emulator-based account.

6.4.1 Commitment as Robust Inclination

We saw that one usage of 'commitment', that of Pepper, might be reflected by the strength of the evidence that an animal takes itself to be acting on, or the intensity of the desire in question. This might lead to the suggestion that the degree of commitment simply corresponds to the degree of inclination. An animal that is more thirsty may have a stronger inclination to find water, and might also be said to be more committed to the goal of finding water. But this does not reflect Bratman's understanding of commitment. For Bratman, the whole point of a committed intention is that it can persist and control your behavior even after the inclination to perform the action has waned, or after competing inclinations have become present. For Bratman, commitment can remain stable *in spite of* what happens to your inclinations. This is reflected in others' accounts of committal desires, committal beliefs, and

[113] Though O'Shaughnessy (1980, p. 546) distinguishes between "cognitive commitments," which result from a decision whether a proposition is true, and "practical commitments," which result from a decision whether to do something.

committal preferences. Commitment can override the inclinational versions of such states, allowing for greater control and coordination of one's reasoning and activities.[114]

But at the same time, commitment and inclination are not completely independent of one another. Commitment raises the threshold that must be met for inclinations to interfere, but strong enough inclinations can overcome commitments. Further, the fact that one has committed to an intention, belief, desire, or preference, may not imply a very high threshold. We have all observed cases of a person's commitment being very brittle; sometimes such people are characterized as "gullible," "influenceable," "fickle," or "capricious." Sometimes the degree of commitment may be high, but the person's effectiveness of self-control may be more compromised (i.e., they thoroughly identify with their commitments but are weak-willed). Here again we see the distinction, discussed in Section 4.5, between degree of *effectiveness of* control and *degree of* control.

6.4.2 Commitment as Higher-Order Inclination

Another way that committal psychological states are often characterized is as being second-order versions of inclinational states (e.g., Ryle, 1949, p. 97). For example, committal desires might be understood as the (inclinational) desire to have a certain (inclinational) desire. I may have a committed desire to quit smoking cigarettes but lack the inclinational desire to do so (i.e., I am still inclined to smoke cigarettes). A natural way to understand my committal desire, on this view, would be that I have an (inclinational) desire to be inclined to quit smoking. Something like this thought lies behind Frankfurt's (1971) earlier "hierarchical" theory of willing and commitment.

[114] Chang (2013, p. 84) provides another argument against treating commitment as a special kind of inclination: whereas you can decide to take on a commitment, you cannot decide to take on an inclination.

This approach has been criticized, however, by Watson (1975), Bratman (2000), and Frankfurt himself (1988). Frankfurt acknowledges Watson's point that merely having a second-order desire does not by itself confer a normative force or stability distinct from first-order desires to rise beyond the status of an inclination:

> The mere fact that one desire occupies a higher level than another in the hierarchy seems plainly insufficient to endow it with greater authority or with any constitutive legitimacy. … Gary Watson has formulated the issue succinctly: "Since second-order volitions are themselves simply desires, to add them to the context of conflict is just to increase the number of contenders; it is not to give a special place to any of those in contention." (Frankfurt, 1988, p. 166; Watson, 1975, p. 218)

Also, the fact that a desire (or belief, or intention) is second-order would not by itself explain other features that Bratman argued are essential to commitments, such as the fact that commitments are resistant to reconsideration and tend to have a greater inertia than inclinations.

6.4.3 Commitment as Evaluative Judgment with Independent Motivating Force

Watson (1975) instead argues that commitments should be construed, not along the lines of desires or intentions, but instead as evaluative judgments. When one commits to quitting smoking, for example, this means that the person has decided that it would be best to quit smoking. In order to explain what Bratman referred to as the "volitional dimension" of commitment, Watson argues that evaluative judgments have a motivational force of their own, independent of inclinations.[115] However, this theory has the implausible consequence that it would not be possible to commit to a plan that one does not believe would be the best thing to

[115] This proposal also of course raises the issue of motivational "internalism" versus "externalism" (Björnsson, Strandberg, Olinder, Eriksson, & Björklund, 2015; or in Staude's 1986 terminology, "cognitivism" versus "conativism"), a debate that is unfortunately outside the scope of this chapter.

do.[116] For example, it would not be possible for someone to decide that it would be best if they did not drive while under the influence of alcohol, but then act on a plan to go to a bar and drive home after drinking. Cases like these provide reason for thinking that one can simultaneously commit to a belief about what would be best to do, and also commit to following a plan that is in conflict with the belief. Such cases also bolster the case I have been making that different kinds of psychological states (e.g., beliefs, intentions, or desires) can involve commitment, and such states will not be functionally equivalent.

6.4.4 Commitment as Personal Identification with Inclinations

After abandoning his earlier hierarchical theory of commitment, Frankfurt adopted a view of commitment based on the idea of personal identification. Frankfurt's way of spelling this out is in terms of "wholeheartedness," which refers to the degree to which "the person's preferences concerning what he wants are … fully integrated" (1988, p. 165). An initial prima facie objection to this proposal might be that wholeheartedness does not seem necessary for commitment. Suppose person A and person B are committed to quitting smoking cigarettes. The fact that person B has stronger cravings than person A does not imply that person B is less committed than person A. Frankfurt anticipates this objection, writing that:

> When someone identifies himself with one rather than with another of his own desires, the result is not necessarily to eliminate the conflict between those desires, or even to reduce its severity, but to alter its nature. Suppose that a person with two conflicting desires identifies with one rather than with the other. This *might* cause the other – the desire with which the person does not identify – to become substantially weaker than it was, or to disappear altogether. But it need not. Quite possibly, the conflict between the two desires will remain as virulent as before. What the person's commitment to the one eliminates is not the conflict between it and the other. It eliminates the conflict *within the person* as to which of these desires he prefers to be his motive. The conflict between the *desires* is in this way transformed into a conflict between *one* of them and the

[116] Similarly, Klein, Molloy, and Cooper argue that "Being committed to a target is distinct from the summary judgment of how favorable (or unfavorable) one views that target. Indeed, one need not have a favorable view of a target to be committed to that target" (2009, p. 9).

> *person* who has identified himself with its rival. That person is no longer uncertain which side he is on, in the conflict between the two desires, and the persistence of this conflict need not subvert or diminish the wholeheartedness of his commitment to the desire with which he identifies. (1988, p. 172)

But now a new problem arises. In what sense is it the "person" that has now become "identified" with the desire? The agent still has both desires. It cannot simply be that the agent has chosen not to have one of the desires. Is the "person" something different from the agent? If so, then it would seem that we are just introducing a homunculus to explain the commitment. We are then led to ask what is involved in the *homunculus* identifying with the desire, and an infinite regress ensues.[117]

Another way of fleshing this out is to say that what one commits to will be what one believes to be most consistent with one's self-conception. But since this conception assimilates committal desires or intentions to beliefs, it faces similar objections to the evaluative judgment approach discussed above. A still further approach is to argue that one's commitments will be those psychological states that match best with one's stable character. The main problem with such virtue-based approaches is that they are not able to account for cases in which a person makes a commitment that is uncharacteristic of them: for example, a violent gang member who decides one day to make a lifestyle change.

Ultimately, I believe that there is an important kernel of truth in the identification theory of commitment. But this is because the agent creates a certain kind of *representation* of itself as constrained by the commitment, and this is a special kind of representation (not merely an evaluative belief or second-order inclination) that can directly influence the agent's behavior. I

[117] Klein, Molloy, and Cooper also argue against understanding commitment in terms of identification:

> Although commitment and identification are both aspects of attachment (Meyer et al., 2006) and are often highly related (Riketta, 2005), identification is both deficient and contaminated as a construct definition of commitment. That is, there are unique aspects of commitment not captured by identification and characteristics of identification that are not part of commitment. (2009, p. 13)

will elaborate on this in Section 6.5. First, it is important to look at another account that I believe is also partially correct.

6.4.5 Frankish on the Relation Between Commitment and Inclination

Frankish (2004, 2016) endorses a distinction between two kinds of beliefs—"basic beliefs" and "superbeliefs"—and two corresponding kinds of desires—"basic desires" and "superdesires." The "super" variety (which he refers to, in general, as "supermental states") share the traits of being "conscious, apt to be occurrently activated, active, flat-out, and frequently language-involving"; the "basic" variety instead have "the opposite properties" (Frankish, 2004, p. 24). In his (2016), he also endorses a distinction between basic and supermental intentions. For Frankish, supermental states, unlike basic folk psychological states, involve commitment in Bratman's sense: they can influence our reasoning and our motivation without the mediation of other people or changes in the environment.

However, Frankish does not consider commitments to be capable of exerting these influences independently of inclinations (i.e., what he calls "basic" beliefs, desires, and intentions). Frankish argues, in fact, that the relation between supermental and basic states is one of *realization*. When one commits to a belief or an intention, one thereby acquires a behavioral disposition in virtue of taking on a constellation of basic-level psychological states:

> a behavioural commitment can be thought of as a kind of disposition – a disposition with a particular sort of basis. If one is committed to A-ing, then one will be disposed to A precisely because one believes oneself to be committed to A-ing and desires to honour this commitment – or, if belief and desire are graded, because one attaches a high probability to the proposition that one is committed to A-ing and a high desirability to honouring the commitment. And, as before, we can think of the commitment as realized in those states. (2004, p. 73)

He understands commitment as a kind of self-imposed cognitive and conative *policy* and writes that "policy-related action is *reflexively motivated*: having a policy of A-ing involves being disposed to A because one believes oneself to have a policy of A-ing and wants to adhere to it"

(2004, p. 109), where the latter belief and desire are basic-level, by which he means they are themselves not only non-committal but also implicit and unconscious. He also refers to the relation as one of *implementation*: "the supermind is implemented, not in the hardware of the brain, but in basic-level intentional states and actions" (2004, p. 7). Frankish also wants to retain an important role of commitments in coordination and planning, however, agreeing with Dennett that committal states "can outlast the beliefs and desires that originally prompted their formation" (Frankish, 2004, p. 73). Thus, while committal states are synchronically "motivated by" (Frankish, 2004, p. 80) what I am calling inclinational states, they do not require the inclinational states involved in their creation to persist.

So far, Frankish's view may seem like a close cousin of Frankfurt's (1971) view. For Frankish, being committed to going to the dentist on Tuesday means wanting to adhere to a policy of going to the dentist on Tuesday and believing that one has taken this on as a policy; this seems suspiciously close to saying (as the 1971 Frankfurt would) that one wants going to the dentist on Tuesday to be one's "will," i.e., to be one's effective inclination. However, Frankish denies that his is such a hierarchical or higher-order type of view:

> Suppose I consciously judge that I need to talk to my bank manager and consciously decide to go to the bank in the morning. Then, these explicit mental states could be cited in explanation of my subsequently going to the bank. However, the conscious decision will have become effective in virtue of implicit mental states, including a belief that I am committed to going to the bank and a desire to execute my commitments, and these implicit states could also be cited in explanation of the action. Since these implicit beliefs and desires concern my premising commitments I shall refer to them as metacognitive states. (Note that 'metacognitive' here does not mean higher-order. The implicit beliefs and desires in question are not about other implicit beliefs and desires but about the premising policies that constitute explicit beliefs and desires.) (2016, p. 38)

Presumably this is because the dispositions that constitute commitment require a complex range of basic beliefs and desires that goes beyond those that would be cited in a higher-order view like Frankfurt's:

> [Adopting a superbelief that] p or deciding to pursue goal x involves adopting a policy of: (1) bearing in mind that one has adopted p as a premise or x as a goal, and looking out for problems and inquiries to which it is relevant; (2) taking p or x

> as input to conscious intentional inference, in conjunction with other premises and goals one has adopted; and (3) acting upon the results of these calculations – adopting any derived propositions and goals as further premises and goals, and performing, or forming intentions to perform, any dictated actions. (2004, p. 97)

If I understand Frankish correctly, then, the basic-level, realizer beliefs and desires will have items such as those listed in this last quotation ("premising policies") as their direct contents (and in this sense, will be "metacognitive"), not "other implicit beliefs and desires" as on a Frankfurt-style view (which would result in Frankish's commitments as being not merely metacognitive, but identical to "second-order" inclinations). This is why the relation between committal and inclinational states is one of realization, not identity:

> Talk of realization is, I think, appropriate here. A high-level behavioural disposition (say, the disposition to save money) exists in virtue of a set of underlying partial beliefs and desires which are, given a normal cognitive background, logically sufficient for it. (Frankish, 2004, p. 65)

In response to Frankish's proposals, it's not clear to me why the "realizers" of committal states should themselves be considered as a form of what he calls "intentional states" (2016, p. 38), i.e., beliefs, desires, or intentions. The kinds of dispositions listed in the above quotation from (2004), p. 97, might occur simply in virtue of internal control system states and processes of the kind I discussed in Chapter 3 of this dissertation, which may have sensory states, control parameters, behaviors, goals, etc. as contents, but without those contents satisfying the criteria I discussed in Chapter 4 (i.e., the kinds of contents beliefs and desires have). Frankish additionally writes that "I do not mean to deny the existence of a level of sub-personal psychology underlying the folk-psychological levels (we might call it a 'sub-mind')" (2004, p. 9); he does not seem to have provided sufficient reason for thinking that states and processes at the "sub-mind" level should not be considered the direct realizers of committal states.

Part of the attractiveness of the idea that there ultimately must be desires or inclinations at the root of all "higher-level" processes like committal intentions, decision-making, etc. may

stem from a further ambiguity of terms like ‘desire’ and ‘inclination’. As a characteristic example exhibiting this ambiguity, consider the following two quotations:

> A brief word on desire. When action occurs, it is in the final analysis this phenomenon that underlies all of the workings of the act-generative mental machinery. For deciding, intending, striving, willing, and choosing, all require a foundation in desire. (O’Shaughnessy, 1980, p. 541)

Twelve pages later in the same volume, O’Shaughnessy writes:

> … is it, after all, credible that a resolve should be something as uncommitted as a mere desire? No matter in what relations we suppose such a desire to stand, those same relations cannot manage to inject into its heart that element of commitment that is so central to the intention. Desire precisely is no form of commitment to its own expression. (1980, p. 553)

Sometimes the word ‘desire’ is used (as in the first passage) to stand in for conative or motivational states of *any* kind; this corresponds to Davidson’s (1963) broad use of the phrase ‘pro attitude’. In this sense, even Bratman would agree that an intention involves “desire”; this is exactly what he referred to as the “volitional dimension” of intention. Similarly, Ryle (1949) used ‘inclination’ in this broad sense to include any kind of motive.

However, such usages of ‘desire’ and ‘inclination’ cannot be operative in passages such as the second one from O’Shaughnessy. In passages like these, ‘desire’ and ‘inclination’ are limited to those motivations that do not result from deliberation or decision but instead “stem from a source external to reason or will” (Schapiro, 2009, p. 233). No philosopher has achieved more mileage from this ambiguity than David Hume, and I believe that to this day, many philosophers like Frankish are tempted by it to analyze even committal states into desires of the non-committal kind.

A second problem with Frankish’s position is that there seems to be a tension between his view that committal states are realized by inclinational states, on the one hand, and his agreement with Dennett (and Bratman) that committal states “can outlast the beliefs and desires that originally prompted their formation” (Frankish, 2004, p. 73). What sort of inclinations might be the ones that realize the outlasting committal states, if not the very same

ones that originally prompted their formation? I believe that Frankish's account may correctly describe the kind of constellation of inclinations that typically accompanies the *formation* of a commitment, but once the commitment has been formed, considerations such as those raised by Bratman make it clear that such underlying inclinations do not need to be sustained for the commitment to persist over time; this would undermine the crucial role of commitment in planning and coordination.

Finally, Frankish's theory is designed to map the committal/non-committal distinction onto other distinctions that are subsumed by "dual-process" theories of reasoning (e.g., Evans & Over, 1996; Stanovich, 1999). Here, recall that Frankish characterizes supermental, but not basic, states as "conscious, apt to be occurrently activated, active, flat-out, and frequently language-involving" (2004, p. 24). However, many of the dimensions along which Frankish divides types of mental states are increasingly coming under criticism in psychology and neuroscience literature (Mugg, 2016, 2018). In his review of Frankish's book, Toribio writes:

> Are all conscious beliefs binary? Are all non-conscious beliefs partial? Is a Bayesian model only applicable to non-conscious reasoning? The division here seems to be too sharp, easily leading to counter-examples, and although Frankish acknowledges that counter-examples would be unavoidable, the alignment of conscious and flat-out beliefs under a classical model of reasoning seems to be just wrong. (2007, p. 140)

Mugg (2018) argues that instead of positing two distinct "minds" or two "reasoning systems," as Frankish does, we need an account of how states with different characteristics interact in a single reasoning system to produce the kind of motivational and cognitive phenomena that correspond to the continuously variable degrees of cognitive control, working memory usage, and adherence to norms of reasoning and planning. In the following section, I provide some suggestions toward a novel account of commitment that meets such criteria.

6.5 An Emulator-Based Hypothesis about the Nature of Commitment

In order to account for self-control through internal commitments, we must go beyond inclinational agency. Something further needs to be built into the agent architecture beyond what inclinational agency includes. First, consider what goes on in the human mind while deliberating about possible courses of action. As described in the previous chapter, an inclinational agent chooses between candidate actions by considering consequences, states of its environmental models that are predicted to result from candidate actions, and compares them based on its preferences about these environmental states. Humans are capable of more than this, however. When deliberating about courses of action, we consider not only the predicted consequences of possible actions, but also existing plans that may conflict. Now a question may be raised here: wouldn't the agent's plans be included in the predictive model? After all, the agent itself is part of its own world, part of its environment; its predicted behaviors would also be taken into account by a "rational" inclinational agent.

This question misses something important: a plan that an agent has is not a *prediction* of what it will do in the future: forming a plan is not merely a process of self-prediction. But while a plan is not a prediction, it is yet something *to be taken into account* when deciding what to do. The total set of plans that an agent has constitute a set of constraints to be taken into account in further deliberation; specifically, a set of constraints *about the agent's future behavior*. In order to properly deliberate in the manner of a human, then, what the agent needs to use is a non-predictive representation of itself, of its own future behavior, that can be updated as its plans continue to be formed. In other words, what the agent needs is a *non-predictive prospective model of its own inclinational self*. I will say that it is a *prescriptive* model

rather than a *predictive* model,[118] although the normativity involved in such prescription is not of the moral variety; it is the minimal normativity of practical rationality: the set of constraints that the agent prescribes to itself in order not to violate its own practical commitments. There is a very simple way to think about the set of constraints constituted by the model: they simply *are* the agent's practical commitments. They constitute the commitments.

Now the constraints of the agent's model of its own inclinational self can only function in the role of *internal commitments* if this model actually has some causal influence over the agent's behavior. Here Bratman would say that we need to account, not only for the reason-centered dimension of commitment, but the *volitional* dimension as well. To capture the volitional dimension of commitment and thereby enable it to function in the role of an internal commitment, the prescriptive model, which could also be called a *committal model*, must be dynamically coupled to the agent's capacity to generate motor impulses.

We already know that the brain makes use of internal models that are causally coupled to one's motor capacities, including at least internal models of the bodily self. Drawing from the work of Mitsuo Kawato and Daniel Wolpert among others, Grush (2004) develops a theory of mental representation based on evidence of the use of neural sensorimotor emulators in motor control. In order for motor control to be as smooth as possible, an emulator in the brain models how one's proprioceptive sensory inputs will change in response to movement. Feedback from the emulator about how movements will affect one's body and the environment can be generated much more rapidly than from sensory signals, so this helps to explain why we are able to engage in such smoothly controlled, precise movements. Grush explains as follows:

[118] Szpunar, Shrikanth, and Schacter differentiate between four types of mental prospection: "*simulation* (construction of a detailed mental representation of the future), *prediction* (estimation of the likelihood of, and/or one's reaction to, a particular future outcome), *intention* (the mental act of setting a goal), and *planning* (the identification and organization of steps toward achieving a goal state)" (2018, p. 52). What they call 'intention' comes closest to what I mean by commitment, but their focus is on the *remembering* of intentions, rather than what Bratman called the volitional dimension of intentions; the latter (which I have referred to as *prescriptive prospection*) does not figure in their taxonomy per se.

> The idea is that in addition to simply engaging with the body and environment, the brain constructs neural circuits that act as models of the body and environment. During overt sensorimotor engagement, these models are driven by efference copies in parallel with the body and environment, in order to provide expectations of the sensory feedback, and to enhance and process sensory information. These models can also be run off-line in order to produce imagery, estimate outcomes of different actions, and evaluate and develop motor plans. (2004, p. 377)

See Figure 6.1. In the case of motor control, the “Plant” represents the musculoskeletal system. The brain generates efferent signals (motor commands) that are sent to the body, but copies of these signals are also sent to circuits in the brain that emulate the musculoskeletal system, generating the ‘mock’ proprioceptive feedback that would result from the movements. The main advantage of using emulated proprioceptive feedback is that it can be generated more quickly than real proprioceptive feedback.

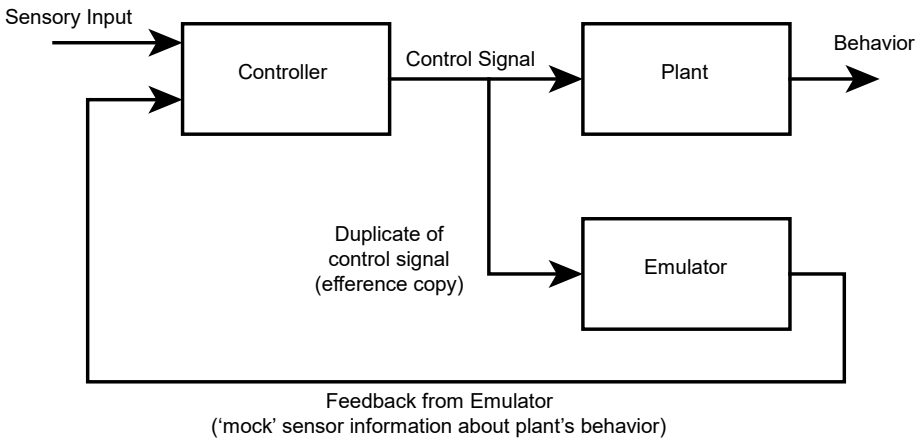

Figure 6.1: Control that relies on feedback received from a forward model (emulator) of the plant (adapted from Grush, 2004, p. 379).

By running “off-line”, Grush means that the control signal is sent only to the emulator, but not to the plant, in Figure 6.1. This allows the motor centers to anticipate how, for example, the position of my arm will hypothetically change if a certain motor signal is produced. As Grush explains, these skills actually require two models: a model of the object itself, or perhaps of myself situated in an environment, and a model of what the resulting sensory inputs would

be given the state of the object or environment (2004, p. 388; both of these would be contained within the “Emulator” box of Figure 6.1).

Grush’s theory is mostly concerned with motor imagery, but he also argues that it can explain how we are able to generate visual imagery, for example, in our imagination. My ability to mentally rotate a cube, for example, may involve an articulated mental model that models a number of parameters of the cube, such as its degree of rotation on three axes, its color, etc. In addition, it would involve a model of how the cube would look from a given perspective and under certain environmental conditions (lighting, etc.), which would take the former model’s parameters (as well as its location and orientation relative to my sensory organs, etc.) as inputs and generate sensory content as output.

The literature on mental and neural models generally treats them as predictive in nature. But what would a *prescriptive* model look like in the brain? The functional roles postulated by Bratman and others for committal states generally fall under what neuroscientists refer to as “executive functions” (e.g., Fuster, 2015, pp. 389–401) or “cognitive control” (e.g., Miller & Cohen, 2001). According to Fuster, the adoption of a plan or distal goal of action involves the updating of a special memory, which he refers to as *executive memory*:

> the prefrontal cortex is the depository of executive memory networks, that is, networks that represent past actions, future actions, or both. It is impossible to construe an executive function without postulating a subjacent executive memory network constituting the neural substrate on which the function will take place. That same network, orderly and timely activated, will be used in attention, in working memory, in planning, and so on. In a word, at a given time, the network will cease to be only representational and will also become operational to serve any or all of those executive functions. (2015, p. 197)

When future actions are stored in executive memory, one is not simply remembering that it is possible or probable that one will take a certain action in the future; by residing in executive memory the future action is *scheduled* to happen, and will be executed (i.e., “operational”) when the time arrives. Based on a wide range of findings that are reviewed in Fuster’s book (2015), Fuster further concludes that

> The abstract schema of the plan would be represented in higher prefrontal (perhaps frontopolar) regions and its more concrete elements of action in lower, premotor, and motor levels. Some of these concrete elements of action may not be represented in the cortex at all but, instead, in the lower levels of the executive hierarchy, such as the cerebellum and the basal ganglia. All in all, the plan is for the organism a way of imagining or creating the future by means of a new or reconstituted neural network. That network, like those that serve the other executive functions, can be appropriately considered a "memory of the future." (2015, p. 391; see also Fuster, 2013, Chapter 5)

Fuster claims that it is difficult to speculate based on current research what kinds of representations are involved in these functions:

> New plans are thoroughly anchored in established executive memory. A new plan is a rearrangement of that memory with a new set of objectives, a new order, a new timetable, and possibly a new ultimate goal. In any case, that plan is essentially based on old experience of prior actions. In a way similar to the way a new perceptual memory is formed in posterior cortex on a base of old memory (phyletic, episodic, semantic, or other), so is the prospective memory of a new plan formed in frontal cortex on a base of established executive memory. By current means, we have no way of knowing how that plan is represented in frontal cortex, least of all how its attributes of time and order are represented. (2015, p. 390)

Pezzulo (2012) suggests that such long-term executive plans may be represented and updated by means of emulators, in a similar way to how emulators enable sensorimotor imagination and coordination. Pezzulo argues that such emulators could represent distal (committal) goals as long-term expectations about future events, and that this can support a view of executive function as an extension of Friston's (2010) "Active Inference" hierarchical theory of brain function.[119] Friston describes action as driven by the minimization of *average surprise*. Here, surprise is being conceived of as a measure of unexpectedness. An input may be unexpected if the agent had not correctly modeled that input as a probabilistic consequence of its causes. If the agent's sensory systems model the environment perfectly, then the agent's average surprise becomes zero, and the recognition density is maximized. The recognition density of the agent is the degree to which it has modeled the causes of its sensory states in a Bayes-optimal fashion. In information theory, a lower average surprise and a greater

[119] Metzinger (2017) makes a similar proposal. See also Pezzulo, Rigoli, and Friston (2018).

recognition density means that there is less *free energy* in the system; this is also expressed by saying that the entropy of the system is lower. According to Friston's free energy formulation of action, then, we can understand perception and action in terms of active inference, which

> can be formulated as a minimization of free-energy; where free-energy bounds the surprise inherent in sensory data, under a model of how those data were caused. This leads to the free-energy principle, which says that everything in the brain should change to minimize free-energy. (Friston et al., 2009, p. 2)

The free energy formulation of action attempts to explain the action of organisms as driven by the reduction of entropy (and thereby the maintenance of self-organization) as well as the minimization of surprise (2010, p. 1). Pezzulo's extension to this framework then proposes that

> cognitive control consists in a *nesting* of optimizations (i.e., free energy minimization loops) over time; in addition to the usual overt loop of active inference, one (or more) covert loop(s) help optimizing distal goals. (2012)

However, since Friston's approach treats all such processes as predictive; it's not clear how it could distinguish between predictive and prescriptive mental prospection, and therefore distinguish between commitments and mere anticipated events (that are anticipated merely by induction or causal reasoning from earlier events). As Bratman (1987, pp. 19–20) argues, we are often able to form intentions that deviate from expectations about what we are most likely to actually do. I can act on my intention to win the lottery by buying a lottery ticket even though I expect to lose. If my commitments really were a matter of minimizing surprise, I could not form such an intention in the first place (unless I were somehow deluded into believing that my winning the lottery is the most likely outcome, which is likely rare among those who play the lottery).[120]

---

[120] Gershman and Daw (2012) use similar considerations to criticize Friston's framework more broadly, arguing that it does not account for the fact that brain processes must often represent the utility of an event separately from its likelihood:

> Although the free-energy principle appears at the least to be a very useful formulation for exposing the computational parallelism between perceptual and decision problems, the

I believe that we can adopt Pezzulo's idea of executive function representations as implemented by emulators, but we have to combine this with Fuster's concept of executive memory. What I am proposing is the notion of an *executive emulator*: an emulator that emulates, not causal environmental or bodily events, but the generation of the effects of inclinational states (such as actions) by other parts of the brain. Emulation could be the means by which executive memory keeps track of past instances of action generation and execution and can recreate the effects of such processes at a later time in the absence of the perceptual stimuli and motivational drive activation that produced the original actions. Suppose I am executing my plan to go to the beach. Emulator circuits are producing similar effects in my brain (generating motor commands, engaging my navigation skills, making me feel frustrated if an obstruction requires me to take a much longer route, etc.) as if I suddenly felt and was acting on an (inclinational) urge to go to the beach. This can occur even if am not currently inclined to go to the beach, i.e., I may be going there to work (e.g., as a lifeguard or surfing instructor), whereas I may feel like doing something somewhere else instead (e.g., go hiking in the mountains). Instead of calling it an "executive emulator," we might call it a *committal emulator*, since it is able to subserve commitments.

How would such an emulator be incorporated into a control scheme? To see this, go back to Figure 6.1 and imagine that the "Plant" is instead the *inclinational self*, i.e., some set of inclinational circuits (the circuits that underwrite inclinational beliefs, desires, preferences, etc.). The "Controller" will be represented by circuits that are upstream from the inclinational circuits

---

> more radical maneuver of treating them both as literally optimizing a single objective function is harder to swallow. A state's equilibrium likelihood and its utility are, on the classical view, not the same thing; rare events might be either unusually bad (being out of water, for a fish), good (being elected president, for an African American), or indeed neither. (2012, p. 306)

in question.[121] The “Emulator” box will then become a committal model of the inclinational self. The emulator will be able to generate anticipated causal consequences of actions, which can be used for on-line cognitive control or off-line deliberation. But a key difference from the scheme in Figure 6.1 is that the emulator will also generate similar output signals to the inclinational circuits and project to similar areas (e.g., the basal ganglia), so that it can compete with the inclinational circuits. The resulting scheme is depicted in Figure 6.2.

This committal emulator theory can explain how committal and inclinational states can compete with each other both cognitively and motivationally within a single reasoning system. Since committal emulators are emulating the inclinational circuits, their outputs will be encoded in the same way, and can be used directly as inputs to other inclinational circuits or to the same

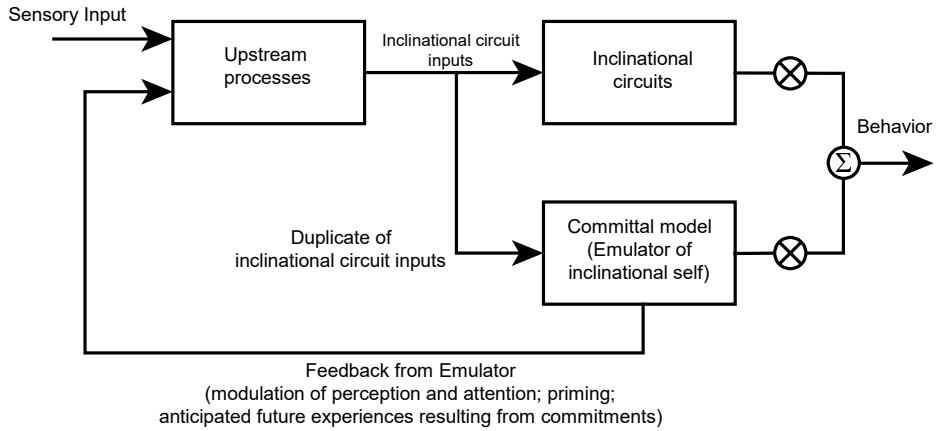

Figure 6.2: Commitment as the emulation of inclination. Note that the committal model does not merely produce anticipated feedback (both present and future), but also directly competes with inclinational circuits to produce motor output. This does not represent a fixed set of components and connections in the brain, but a control scheme that can exist at and spontaneously organize on the fly at many levels of organization. Different committal emulators may become active depending on what the subject is deliberating about or what their current tasks and goals are.

[121] As described by Berntson and Cacioppo (2008), motivational circuits are organized heterarchically, not hierarchically, in the brain, so which circuits are “upstream” from which ones will depend on the functional context.

outputs as inclinational circuits. This could explain the extensive connectivity running directly from the PFC to late motor areas, basal ganglia, and cerebellum.

Similarly, inclinational circuits can input directly to the committal emulators. A committal emulator might be used for more deliberately controlled performance of an action that is stimulated inclinationally, as when I feel a sudden inclination to shoot while playing basketball, but then effortfully try to control the motion in a certain way that I earlier consciously planned (i.e., I shoot because I am inclined to, but then I make sure to shoot in the way that I committally desired to, e.g., following a coach's advice of keeping my shoulders square to the basket). In this way, I can smoothly transition between executive and automatic control.

The use of committal emulators can also help to explain why mentally simulating future actions (Taylor & Pham, 1996) and forming implementation intentions (Gollwitzer, 1999) increases commitment: these actions stimulate off-line emulation of the circuits involved in the action, causing reinforcement of and articulation of the circuits subserving the emulation in executive memory. Further, the fact that humans are able to anticipate future motivational states provides support for the idea that commitment may rely on the emulation of inclinational circuits. As Sharot writes:

> humans are able to take into account future motivational states when making decisions. We go grocery shopping, even if we had recently eaten a large meal, because we can anticipate being hungry in a few hours and are aware that the fridge is empty (although we might purchase less than we eventually need). This ability may be absent in our less sophisticated evolutionary ancestors. (2012, pp. 67–68)

In cases like this, a committal emulator may be running off-line to facilitate a decision of whether to go shopping, and if so, what to buy.

As noted by Mugg (2018), the human brain is not limited to discretely switching in an all-or-nothing way between using automatic and controlled (or "Type 1" and "Type 2") processes for a given task; humans can smoothly vary the amount of working memory and attention being used, and the extent to which the task is performed "consciously" or

“deliberately.” Again, the emulator theory can provide insight. Grush (2004) discusses the fact that sensorimotor neural emulators likely use something that works like a Kalman filter (Kálmán & Bucy, 1961) to continuously compare the actual proprioceptive feedback with the proprioceptive feedback that is modeled by the emulator circuits. The difference is used to make corrections in the operation of the emulator. But the system will also keep track of the reliability of the proprioceptive feedback signal; when it is too noisy or unreliable, the system will weight the modeled proprioceptive feedback from the emulator more heavily so that it has greater influence in regulating motor output (this weighting is called the “Kalman gain”). The Kalman gain therefore determines the extent to which the dynamics of the internal model are relied upon for motor control at a given time rather than the dynamics of the body itself. The Kalman filter then allows motor control to be much smoother and more robust than it would be if only proprioceptive feedback or only emulator feedback were used (Wolpert, Ghahramani, & Jordan, 1995).

Something like Kalman filters could be in operation to bias the amount of influence the executive emulators will have versus subcortical motivational circuits over action selection and motor control (in Figure 6.2, their point of modulation is designated with the symbol ⊗). The neuromodulatory systems could play a key role here, which would make sense given the widespread connectivity of such systems to both cortical and subcortical areas of the brain (Briand et al., 2007). For example, suppose someone is hiking in the wilderness, following a route that they have deliberate decided on days earlier. But the person suddenly sees a dangerous animal charging and “instinctively” starts running away from it, deviating from the planned route. Of course, it is likely that no reconsideration of one’s intention was necessary in this case; the brain’s noradrenergic system likely causes the equivalent of a large shift in the “Kalman gain,” resulting in one’s fight-or-flight inclinational states to take almost total control. Similarly, by stimulating the GABA system, drugs like Ritalin can have an opposite effect on the “Kalman gain,” causing the executive emulator circuits to have more control, and controlling

inclinational impulses. Continuing this analogy, the process that is analogous to monitoring the reliability or noise-level of the proprioceptive feedback is likely far more complex, and the emotions and reward system likely also play key roles.[122] There are likely many different "Kalman gains" corresponding to different kinds of inclinational circuits (and corresponding emulators) that may operate in a more or less "automatic" manner.

Committal emulators could also be used to represent the commitments of other people for use in social interaction and coordination. As Ybarra and Winkielman write, humans engage in

> a diverse range of … intense social interactions, spanning the gamut from forming impressions on dates and performance in job interviews, to discussions of policies with colleagues, consultations of clothing choices with a sensitive spouse, bargaining with a sneaky salesperson, to performing a complex dance, playing a bridge game, or having a diplomatic negotiation. Often, to be effective in the latter, complex and often more consequential type of social interaction a perceiver is required to develop an on-line representation of a dynamic and changing mental-model of another person's beliefs, expectations, emotions, and desires. The perceiver also must be able to problem-solve, inhibit inappropriate behaviors, take-turns, and pursue goals in a distraction-rich environment. (2012, p. 1)

It is possible that executive functions and the ability for these kinds of social interactions evolved together, as the ability to construct and use committal emulators evolved. This would also explain findings that suggest that complex social interactions can facilitate more effective executive functioning (Ybarra and Winkielman, 2012; Ding et al., 2018); such activities might do so by training the brain's ability to generate and use committal emulators.

It would certainly not be possible in the limited space of this chapter to evaluate the committal emulator hypothesis in light of the large body of empirical findings on executive

[122] The hypothesis being developed here is admittedly very sketchy and it is too early for empirical confirmation or disconfirmation. However, Avery and Krichmar claim that empirical findings support the view that neuromodulatory systems "provide a foundation for cognitive function in higher organisms; attention, emotion, goal-directed behavior, and decision-making derive from the interaction between the neuromodulatory systems and brain areas, such as the amygdala, frontal cortex, hippocampus, and sensory cortices" (2017, p. 1). Such findings would mesh well with the view that neuromodulatory systems facilitate the development and updating of executive emulator circuits and modulation of competition between executive emulators and inclinational circuits.

functioning and neuromodulatory systems to date, nor to speculate about specific empirical predictions that would follow from it; these are tasks to be pursued in future work. Instead I will conclude this section by considering some differences between the position here being explored and Frankish's position that was discussed in the previous section. Recall that on Frankish's view, committal states are not identical to inclinational states but are *implemented by* them. For Frankish this is a synchronic relation: a committal state exists at a given moment in virtue of the inclinational states that implement it at that moment. This is not the case on the committal emulator hypothesis: committal states are implemented by emulator circuits, which model inclinational circuits. But the inclinational circuits they model do not need to be actively engaged for a commitment to influence one's behavior or cognition (just as I act on my commitment to go to the dentist even though I don't really feel like going), and may no longer even exist in the same form as when the emulator circuit was first formed. Emulator circuits could also be modulated without the corresponding inclinational circuit being modulated in that way; they may be subserved by different parts of the brain. This better accommodates Dennett's and Bratman's point that, as Frankish puts it, committal states "can outlast the beliefs and desires that originally prompted their formation" (2004, p. 73).

Like Frankish's view, I believe that the committal emulator picture can support both committal and inclinational versions of desires, intentions, and beliefs. Executive emulators could emulate not only inclinational desire or inclinational intention circuits, but also inclinational circuits that register perceptual information or the presence/absence of states of affairs (to yield committal beliefs). Since it is the functional (input and output) characteristics that are emulated rather than the internal workings, the corresponding imagery or perceptual data would not need to be emulated. This is likely the reason why executive emulators lend themselves more naturally to articulation of their contents in linguistic form. Unlike on Frankish's view, however, I don't think that conceptual or linguistic articulation is essential to committal states or the executive emulators that underlie them.

Similarly, I don't think that committal states are necessarily more "conscious" than inclinational states. What does seem to be true about committal emulators is that their use requires working memory and attentional resources and can result in "ego depletion" (Baumeister, Schmeichel, & Vohs, 2007). But the work of Bargh (e.g., Huang & Bargh, 2014), among others, seems to demonstrate that goal commitments can be formed, sustained, modified, and can control behavior without conscious awareness, conscious effort, or the "feeling of agency." Finally, though committal states often seem to work in an all-or-nothing manner (e.g., flat-out as opposed to Bayesian belief), this likely does not reflect the nature of committal states in general but instead the way they often interact with other states. Inclinational states work in a partial way because they are frequently in competition with other conflicting inclinational states—an unlimited number of inclinational circuits can be activated at a given time. But due to the fact that committal states generally require scarce working memory and attentional resources, far fewer of them can be actively engaged at the same time. Additionally, nothing stops a person from committing to a partial belief. After thinking it over for a while, one might make up one's mind that there is a pretty good chance that the person they saw earlier that day was someone they went to elementary school with. Having committed to this (partial) belief, one may stick to it even after one's memory of the experience that first triggered the inclinational belief has faded (and with it, the inclinational belief).

6.6 Conclusion: Internal Perspectivalism about Committal Agency

All committal agents are agents. All agents are control systems (not mere regulators; see Section 4.4.1). All control systems are observer-worker systems. Because committal agents are observer-worker systems, they (at any moment) have a perspective on how to behave given the current detected situation. Because they are control systems, they have a perspective on what kind of influence to have on control parameters given the current sensory

(input) parameters. Because they are agents (i.e., at least have the capacity for inclinational agency), they have preferences, which requires a perspective on whether things that are potentially either under their control or not under their control are in fact under their control at a given time. Continuing this trend, we might ask if committal agency is associated with a distinct type of perspective. Here, I believe that the following quotation from Bratman is especially relevant:

> Of course, there is a sense in which when I act, I act at a particular time; but in acting I do not see myself, the agent of the act, as simply a time-slice agent. I see my action at that time as the action of the same agent as he who has acted in the past and (it is to be hoped) will act in the future. In this respect I differ importantly from those nonhuman agents who do not have the resources to understand their own agency as temporally extended. (2000, p. 43)

I believe that Bratman's insight here, that what is distinctive about committal agency (which he calls "planning agency") is that *from the agent's perspective*, the agent's own agency is temporally extended, is fundamentally correct. The agent's own self, *as agent*, becomes not just the subject but the object *of its own agency*: it becomes one of the things that has been, and will be, either under its control or not under its control in various ways at various times. It doesn't just see its own body or its cognitive capacities as tools for its own use: it sees its own *agency* as a tool for its own use. And it can subject itself to its own past commands, as well as subordinate its future self to its current self, by means of internal operations. The emulator theory of commitment provides a way of fleshing this out mechanistically.

References

Allen, N. J. (2016). Commitment as a Multidimensional Construct. In J. P. Meyer (Ed.), *Handbook of Employee Commitment* (pp. 28–42). Cheltenham, UK: Elgar.

Andreou, C. (2007). There Are Preferences and Then There Are Preferences. In B. Montero & M. D. White (Eds.), *Economics and the Mind* (pp. 115–126). London: Routledge.

Anscombe, G. E. M. (1963). *Intention* (2nd Ed.). Cambridge, MA: Harvard University Press.

Arbib, M. (1981). Perceptual Structures and Distributed Motor Control. In V. B. Brooks (Ed.), *Handbook of Physiology, Section 1: The Nervous System* (Vol. 2). Bethesda, MD: American Physiological Association.

Arkin, R. C. (1998). *Behavior-Based Robotics*. Cambridge, MA: MIT Press.

Arnellos, A., Bruni, L. E., El-Hani, C. N., & Collier, J. (2012). Anticipatory Functions, Digital-Analog Forms and Biosemiotics: Integrating the Tools to Model Information and Normativity in Autonomous Biological Agents. *Biosemiotics*, *5*(3), 331–367.

Ashby, W. R. (1956). *An Introduction to Cybernetics.* London: Chapman and Hall.

Ashby, W. R. (1962). Principles of the Self-Organizing System. In H. von Foerster & G. W. Zopf (Eds.), *Principles of Self-Organization* (pp. 255–278). New York: Pergamon.

Atkins, P. W. (1984). *The Second Law: Energy, Chaos, and Form*. New York: Scientific American.

Avery, M. C., & Krichmar, J. L. (2017). Neuromodulatory Systems and Their Interactions: A Review of Models, Theories, and Experiments. *Frontiers in Neural Circuits*, *11*(108).

Baier, A. (1979). Mind and Change of Mind. *Midwest Studies in Philosophy, 4*(1), 157–176.

Bandura, A. (2001). Social Cognitive Theory: An Agentic Perspective. *Annual Review of Psychology*, *52*(1), 1–26.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals. *Journal of Personality and Social Psychology*, *81*(6), 1014–1027.

Baron, J. (2007). *Thinking and Deciding* (4th Ed.). Cambridge: Cambridge University Press.

Barwise, J., & Perry, J. (1983). *Situations and Attitudes*. Cambridge, MA: Bradford.

Baumeister, R. F., Schmeichel, B. J., & Vohs, K. D. (2007). Self-Regulation and the Executive Function: The Self as Controlling Agent. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social Psychology: Handbook of Basic Principles* (2nd Ed., pp. 516–539). New York: Guilford.

Beatty, J. (1984). Chance and Natural Selection. *Philosophy of Science, 51*, 183–211.

Bechtel, W. (1986). Teleological Functional Analyses and the Hierarchical Organization of Nature. In N. Rescher (Ed.), *Current Issues in Teleology* (pp. 26–48). Landham, MD: University Press of America.

Bechtel, W. (2006). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge: Cambridge University Press.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Routledge.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A Mechanist Alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421–441.

Bedau, M. (1992). Goal-Directed Systems and the Good. *Monist, 75*(1), 34–51.

Bennett, J. (1976). *Linguistic Behaviour*. Cambridge: Cambridge University Press.

Berntson, G. G., & Cacioppo, J. T. (2008). The Neuroevolution of Motivation. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of motivation science* (pp. 188–200). New York: Guilford.

Berridge, K. C. (2018). Evolving Concepts of Emotion and Motivation. *Frontiers in Psychology, 9*, 1647. doi:10.3389/fpsyg.2018.01647

Bickhard, M. H. (2000). Autonomy, Function, and Representation. *Communication and Cognition: Artificial Intelligence, 17*(3–4), 111–131.

Bickhard, M. H., & Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*. Amsterdam: North-Holland.

Bigelow, J., & Pargetter, R. (1987). Functions. *Journal of Philosophy, 84*(4), 181–196.

Bishop, J. (1989). *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge: Cambridge University Press.

Bishop, M. A. (2010). Why the Generality Problem is Everybody's Problem. *Philosophical Studies, 151*, 285–298.

Bishop, R. C. (2012). Fluid Convection, Constraint and Causation. *Interface Focus, 2*, 4–12.

Björnsson, G., Strandberg, C., Olinder, R. F., Eriksson, J., & Björklund, F. (2015). Motivational Internalism: Contemporary Debates. In G. Björnsson, C. Strandberg, R. F. Olinder, J. Eriksson, & F. Björklund (Eds.), *Motivational Internalism* (pp. 1–20). Oxford: Oxford University Press.

Bradburn, N. M., Cartwright, N., & Fuller, J. (2017). A Theory of Measurement. In L. McClimans (Ed.), *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation* (pp. 73–88). London: Rowman & Littlefield.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology.* Cambridge, MA: MIT Press.

Brandom, R. B. (1994). *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Brandom, R. B. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (1992). Practical Reasoning and Acceptance in a Context. *Mind, 101*, 1–14.

Bratman, M. E. (1996). Identification, Decision, and Treating as a Reason. *Philosophical Topics, 24*(2), 1–18.

Bratman, M. E. (2000). Reflection, Planning, and Temporally Extended Agency. *Philosophical Review*, *109*(1), 35–61.

Braver, T. S., Krug, M. K., Chiew, K. S., Kool, W., Westbrook, J. A., Clement, N. J., … Somerville, L. H. (2014). Mechanisms of Motivation-Cognition Interaction: Challenges and Opportunities. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 443–472. https://doi.org/10.3758/s13415-014-0300-0

Brehm, J. W., & Cohen, A. R. (1962). *Explorations in Cognitive Dissonance*. New York: Wiley.

Briand, L. A., Gritton, H., Howe, W. M., Young, D. A., & Sarter, M. (2007). Modulators in Concert for Cognition: Modulator Interactions in the Prefrontal Cortex. *Progress in Neurobiology*, *83*(2), 69–91.

Brooks, R. A. (1991). Intelligence without Reason. In J. Mylopoulos & R. Reiter (Eds.), *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 569–595). San Mateo, CA: Morgan Kaufmann.

Brunstein, J. C. (1993). Personal Goals and Subjective Well-Being: A Longitudinal Study. *Journal of Personality and Social Psychology*, *65*(5), 1061–1070.

Burge, T. (2009). Primitive Agency and Natural Norms. *Philosophy and Phenomenological Research, 79*(2), 251–278.

Burge, T. (2010). *Origins of Objectivity*. Oxford: Clarendon.

Campbell, R. (2009). A Process-Based Model for an Interactive Ontology. *Synthese, 166*, 453–477.

Cantor, N., & Sanderson, C. A. (1999). Life Task Participation and Well-Being: The Importance of Taking Part in Daily Life. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well Being: The Foundations of Hedonic Psychology* (pp. 230–243). New York: Russell Sage Foundation.

Carbonelle, J. G. (1982). Where Do Goals Come From? In *Proceedings of the Fourth Annual Conference of the Cognitive Science Society* (pp. 191–194). Austin, TX: Cognitive Science Society.

Carruthers, P. (2007). The Illusion of Conscious Will. *Synthese*, *159*(2), 197–213.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement.* Oxford: Clarendon.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science.* Cambridge: Cambridge University Press.

Cartwright, N., & Pemberton, J. (2013). Aristotelian Powers: Without Them, What Would Modern Science Do? In R. Groff & J. Greco (Eds.), *Powers and Capacities in Philosophy: The New Aristotelianism* (pp. 93–112). New York: Routledge.

Carver, C. S., & Scheier, M. F. (1981). *Attention and Self-Regulation: A Control-Theory Approach to Human Behavior*. New York: Springer-Verlag.

Castelfranchi, C. (1995). Guarantees for Autonomy in Cognitive Agent Architecture. In M. J. Wooldridge & N. R. Jennings (Eds.), *Intelligent Agents* (pp. 56–70). Berlin: Springer-Verlag.

Chang, R. (2013). Commitments, Reason, and the Will. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 8, pp. 74–113). Oxford: Oxford University Press.

Chang, R. (2017). Hard Choices. *Journal of the American Philosophical Association*, *3*(1), 1–21.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: Bradford.

Chopra, A. K., & Singh, M. P. (2013). Agent Communication. In G. Weiss (Ed.), *Multiagent Systems* (2nd Ed., pp. 101–141). Cambridge, MA: MIT Press.

Christensen, W. D. (1999). *An Interactivist-Constructivist Approach to Adaptive Intelligence and Agency* (Dissertation). University of Newcastle, Australia.

Christensen, W. D., & Bickhard, M. H. (2002). The Process Dynamics of Normative Function. *Monist, 85*(1), 3–28.

Cialdini, R. B. (2009). *Influence: Science and Practice* (5th Ed.). Boston: Pearson.

Cohen, L. J. (1986). *The Dialogue of Reason: An Analysis of Analytical Philosophy*. Oxford: Clarendon.

Cohen, L. J. (1992). *An Essay on Belief and Acceptance*. Oxford: Clarendon.

Collier, J. D., & Hooker, C. A. (1999). Complexly Organised Dynamical Systems. *Open Systems and Information Dynamics, 6*, 241–302.

Conee, E., & Feldman, R. (1998). The Generality Problem for Reliabilism. *Philosophical Studies, 89*, 1–29.

Copp, D. (2015). Explaining Normativity. *Proceedings and Addresses of the American Philosophical Association, 89*, 48–73.

Coutlee, C. G., & Huettel, S. A. (2014). Rules, Rewards, and Responsibility: A Reinforcement Learning Approach to Action Control. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 4: Free Will and Moral Responsibility* (pp. 327–334). Cambridge, MA: Bradford.

Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science, 68*(1), 53–74.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon.

Craver, C. F. (2013). Functions and Mechanisms: A Perspectivalist View. In P. Huneman (Ed.), *Functions: Selection and Mechanisms* (pp. 133–158). Dordrecht: Springer.

Craver, C. F., & Bechtel, W. (2006). Mechanism. In S. Sarkar & J. Pfeifer (Eds.), *The Philosophy of Science: An Encyclopedia* (pp. 469–478). New York: Routledge.

Cummins, R. (1975). Functional Analysis. *Journal of Philosophy, 72*(20), 741–765.

Curioni, A., Knoblich, G., & Sebanz, N. (2016). Joint Action in Humans: A Model for Human-Robot Interactions. In A. Goswami & P. Vadakkepat (Eds.), *Humanoid Robotics: A Reference* (pp. 1–19). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7194-9_126-1

Cussins, A. (1990). The Connectionist Construction of Concepts. In M. Boden (Ed.), *The Philosophy of Artificial Intelligence* (pp. 368–440). Oxford: Oxford University Press.

Cussins, A. (1992). Content, Embodiment and Objectivity: The Theory of Cognitive Trails. *Mind, 101*(404), 651–688.

Darden, L. (2008). Thinking Again about Biological Mechanisms. *Philosophy of Science, 75*, 958–969.

Daveney, T. F. (1961). Wanting. *Philosophical Quarterly, 11*(43), 135–144.

Daveney, T. F. (1964). Choosing. Reprinted in M. Brand (Ed.), *The Nature of Human Action* (pp. 82–90). Glenview, IL: Scott, Foresman and Company, 1970.

Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy*, *60*(23), 685–700.

Davidson, D. (1973). Freedom to Act. In T. Honderich (Ed.), *Essays on Freedom of Action* (pp. 137–156). London: Routledge and Kegan Paul.

Davidson, D. (1982). Rational Animals. *Dialectica, 36*, 317–328.

Davidson, D. (1999). The Emergence of Thought. *Erkenntnis*, *51*(1), 7–17.

Davies, P. C. W. (1989). The Physics of Complex Organisation. In B. Goodwin & P. Saunders (Eds.), *Theoretical Biology: Epigenetic and Evolutionary Order for Complex Systems* (pp. 101–111). Edinburgh: Edinburgh University Press.

Davies, P. S. (2001). *Norms of Nature*. Cambridge, MA: Bradford.

Davis, W. A. (1984). The Two Senses of Desire. *Philosophical Studies*, *45*(2), 181–195.

Dayan, P. (2012). Models of Value and Choice. In R. J. Dolan & T. Sharot (Eds.), *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms* (pp. 33–52). London: Academic Press.

de Sousa, R. B. (1971). How to Give a Piece of Your Mind: Or, the Logic of Belief and Assent. *Review of Metaphysics, 25*(1), 52–79.

Deacon, T. W. (2007). Shannon–Boltzmann–Darwin: Redefining Information (Part 1). *Cognitive Semiotics, 1*(Fall 2007), 123–148.

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter.* New York: W. W. Norton.

Dennett, D. C. (1978). How to Change Your Mind. In his *Brainstorms* (pp. 300–309). Montgomery, VT: Bradford.

Dennett, D. C. (1981). True Believers: The Intentional Strategy and Why It Works. Reprinted in his *The Intentional Stance* (pp. 13–35). Cambridge, MA: Bradford, 1987.

Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). Real Patterns. *Journal of Philosophy, 88*(1), 27–51.

Dennett, D. C. (1996). *Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic.

Devlin, K. (1991). *Logic and Information*. Cambridge: Cambridge University Press.

Di Paolo, E. A. (2005). Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenology and the Cognitive Sciences*, *4*(4), 429–452.

Dickinson, A., & Balleine, B. W. (1995). Motivational Control of Instrumental Action. *Current Directions in Psychological Science*, *4*(5), 162–167.

Dickinson, A., & Balleine, B. W. (2000). Causal Cognition and Goal-Directed Action. In C. Heyes & L. Huber (Eds.), *The Evolution of Cognition* (pp. 185–204). Cambridge, MA: Bradford.

Dickinson, A., & Balleine, B. W. (2002). The Role of Learning in the Operation of Motivational Systems. In R. Gallistel (Ed.), *Stevens' Handbook of Experimental Psychology: Learning, Motivation, and Emotion* (3rd. ed., Vol. 3, pp. 497–533). New York: John Wiley & Sons.

Ding, X. P., Heyman, G. D., Sai, L., Yuan, F., Winkielman, P., Fu, G., & Lee, K. (2018). Learning to Deceive Has Cognitive Benefits. *Journal of Experimental Child Psychology*, *176*, 26–38.

Dirac, P. A. M. (1958). *The Principles of Quantum Mechanics* (4th ed.). Oxford: Clarendon.

Dodd, D. (2009). Weakness of Will as Intention-Violation. *European Journal of Philosophy, 17*(1), 45–59.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Drescher, G. (1991). *Made-Up Minds: A Constructivist Approach to Artificial Intelligence.* Cambridge, MA: MIT Press.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: Bradford.

Dretske, F. I. (1986). Aspects of Cognitive Representation. In M. Brand & R. M. Harnish (Eds.), *The Representation of Knowledge and Belief* (pp. 101–115). Tucson: University of Arizona Press.

Dretske, F. I. (1988). *Explaining Behavior: Reasons in a World of Causes.* Cambridge, MA: Bradford.

Dretske, F. I. (1995). *Naturalizing the Mind*. Cambridge, MA: Bradford.

Dretske, F. I. (1999). Machines, Plants and Animals: The Origins of Agency. *Animal Mind, 51*(1), 19–31.

Dupré, J. (2007). *The Constituents of Life*. Amsterdam: Van Gorcum. Reprinted in J. Dupré, *Processes of Life: Essays in the Philosophy of Biology*, Oxford: Oxford University Press, 2012.

Elster, J. (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Emirbayer, M., & Mische, A. (1998). What Is Agency? *American Journal of Sociology*, *103*(4), 962–1023.

Enç, B. (2002). Indeterminacy of Function Attributions. In A. Ariew, R. Cummins, & M. Perlman (Eds.), *Functions: New Essays in the Philosophy of Psychology and Biology* (pp. 291–313). Oxford: Oxford University Press.

Esfeld, M. (2009). The Modal Nature of Structures in Ontic Structural Realism. *International Studies in the Philosophy of Science, 23*(2), 179–194.

Evans, G. (1975). Identity and Predication. *Journal of Philosophy, 72*(13), 343–363.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Clarendon.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and Reasoning.* Hove, UK: Psychology Press.

Falk, W. D. (1948). “Ought” and Motivation. *Proceedings of the Aristotelian Society, 48*, 111–138.

Fischer, J. M. (1994). *The Metaphysics of Free Will: An Essay on Control*. Oxford: Blackwell.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Foot, P. (2001). *Natural Goodness*. Oxford: Clarendon.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy, 68*, 5–20.

Frankfurt, H. G. (1978). The Problem of Action. *American Philosophical Quarterly, 15*(2), 157–162.

Frankfurt, H. G. (1988). Identification and Wholeheartedness. In his *The Importance of What We Care About* (pp. 159–176). Cambridge: Cambridge University Press.

Frankish, K. (2004). *Mind and Supermind*. Cambridge: Cambridge University Press.

Frankish, K. (2016). Playing Double: Implicit Bias, Dual Levels, and Self-Control. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy* (Vol. 1, pp. 23–46). Oxford: Oxford University Press.

Franklin, G. F., Powell, J. D., & Emami-Naeini, A. (2018). *Feedback Control of Dynamic Systems* (8th ed.). Boston: Pearson.

French, S. (2006). Structure as a Weapon of the Realist. *Proceedings of the Aristotelian Society, 106*, 167–185.

French, S. (2014). *The Structure of the World: Metaphysics and Representation.* Oxford: Oxford University Press.

Fridland, E. (forthcoming). *Skill in Action*. Oxford: Oxford University Press.

Frijda, N. (2007). *The Laws of Emotion*. Mahwah, NJ: Lawrence Erlbaum.

Friston, K. J. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J., Daunizeau, J., & Kiebel, S. (2009). Reinforcement Learning or Active Inference? *PLoS One, 4*(7), e6421.

Funkenstein, A. (1986). *Theology and the Scientific Imagination: From the Middle Ages to the Seventeenth Century.* Princeton, NJ: Princeton University Press.

Fuster, J. M. (2013). *The Neuroscience of Freedom and Creativity: Our Predictive Brain*. Cambridge: Cambridge University Press.

Fuster, J. M. (2015). *The Prefrontal Cortex* (5th Ed.). Amsterdam: Academic.

Gallese, V., & Metzinger, T. (2003). Motor Ontology: The Representational Reality of Goals, Actions and Selves. *Philosophical Psychology, 16*(3), 365–388.

Galluzzi, L., Bravo-San Pedro, J. M., & Kroemer, G. (2014). Organelle-Specific Initiation of Cell Death. *Nature Cell Biology, 16*(6), 728–736.

Garson, J. (2013). The Functional Sense of Mechanism. *Philosophy of Science, 80*, 317–333.

Gendler, T. S. (2008). Alief and Belief. *Journal of Philosophy, 105*(10), 634–663.

Gershman, S. J., & Daw, N. D. (2012). Perception, Action, and Utility: The Tangled Skein. In M. I. Rabinovich, K. J. Friston, & P. Varona (Eds.), *Principles of Brain Dynamics: Global State Interactions* (pp. 293–312). Cambridge, MA: MIT Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gilbert, S. F., & Barresi, M. J. F. (2018). *Developmental Biology* (11th Ed.). Sunderland, MA: Sinauer.

Gill, M. L. (1989). *Aristotle on Substance: The Paradox of Unity.* Princeton, NJ: Princeton University Press.

Gillett, C. (2007). Understanding the New Reductionism: The Metaphysics of Science and Compositional Reduction. *Journal of Philosophy, 104*(4), 193–216.

Glennan, S. (1996). Mechanisms and the Nature of Causation. *Erkenntnis, 44*(1), 49–71.

Glennan, S. (2009). Mechanisms. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford Handbook of Causation* (pp. 315–325). Oxford: Oxford University Press.

Glennan, S. (2017). *The New Mechanical Philosophy.* Oxford: Oxford University Press.

Globus, G. G. (1976). Mind, Structure, and Contradiction. In G. G. Globus, G. Maxwell, & I. Savodnik (Eds.), *Consciousness and the Brain: A Scientific and Philosophical Inquiry* (pp. 271–293). New York: Plenum.

Goldman, A. I. (1979). What is Justified Belief? In G. S. Pappas (Ed.), *Justification and Knowledge: New Studies in Epistemology* (pp. 1–23). Dordrecht: Reidel.

Goldstein, H., Poole, C., & Safko, J. (2002). *Classical Mechanics* (3rd ed.). San Francisco: Addison Wesley.

Golledge, R. (2010). Cognitive Maps. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini Encyclopedia of Psychology*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470479216.corpsy0196

Gollwitzer, P. M. (1999). Implementation Intentions: Strong Effects of Simple Plans. *American Psychologist*, *54*(7), 493–503.

Gollwitzer, P. M., & Oettingen, G. (2011). Planning Promotes Goal Striving. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of Self-Regulation: Research, Theory, and Applications* (2nd Ed., pp. 162–185). New York: Guilford.

Goode, R., & Griffiths, P. E. (1995). The Misuse of Sober's Selection for/Selection of Distinction. *Biology and Philosophy, 10*, 99–108.

Goodman, N. (1954). *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press.

Greeno, J. G. (1994). Gibson's Affordances. *Psychological Review, 101*(2), 336–342.

Grush, R. (2000). Self, World, and Space: The Meaning and Mechanisms of Ego- and Allocentric Spatial Representation. *Brain and Mind, 1*, 59–92.

Grush, R. (2004). The Emulation Theory of Representation: Motor Control, Imagery, and Perception. *Behavioral and Brain Sciences*, *27*(3), 377–442.

Grush, R. (2007). Berkeley and the Spatiality of Vision. *Journal of the History of Philosophy, 45*(3), 413–442.

Grush, R. (2009). Space, Time, and Objects. In J. Bickle (Ed.), *The Oxford Handbook of Philosophy and Neuroscience* (pp. 311–345). Oxford: Oxford University Press.

Grush, R., & Springle, A. (forthcoming). Agency, Perception, Space, and Objectivity. *Phenomenology and the Cognitive Sciences.*

Harmon-Jones, E., & Harmon-Jones, C. (2008). Cognitive Dissonance Theory: An Update with a Focus on the Action-Based Model. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of Motivation Science* (pp. 71–83). New York: Guilford.

Harré, H. R., & Madden, E. H. (1975). *Causal Powers: A Theory of Natural Necessity.* Oxford: Basil Blackwell.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: Bradford.

Haugeland, J. (1990). The Intentionality All-Stars. Reprinted in his *Having Thought: Essays in the Metaphysics of Mind* (pp. 127–170). Cambridge, MA: Harvard University Press, 1998.

Hautamäki, A. (1986). *Points of View and Their Logical Analysis*. Helsinki: Philosophical Society of Finland.

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (2012). *Acceptance and Commitment Therapy: The Process and Practice of Mindful Change* (2nd Ed.). New York: Guilford.

Heck, R. G. (2000). Nonconceptual Content and the "Space of Reasons." *Philosophical Review, 109*(4), 483–523.

Heil, J. (2005). Dispositions. *Synthese, 144*, 343–356.

Heil, J. (2012). *The Universe as We Find It.* Oxford: Clarendon.

Henden, E. (2008). What is Self-Control? *Philosophical Psychology, 21*(1), 69–90.

Henry, J. (2004). Metaphysics and the Origins of Modern Science: Descartes and the Importance of Laws of Nature. *Early Science and Medicine, 9*(2), 73–114.

Hinde, R. A. (1999). *Why Gods Persist: A Scientific Approach to Religion*. London: Routledge.

Holland, J. H. (2012). *Signals and Boundaries: Building Blocks for Complex Adaptive Systems.* Cambridge, MA: MIT Press.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Clarendon.

Hommel, B., & Wiers, R. W. (2017). Towards a Unitary Approach to Human Action Control. *Trends in Cognitive Sciences*, *21*(12), 940–949.

Hooker, C. A. (2009). Interaction and Bio-Cognitive Order. *Synthese, 166*, 513–546.

Hooker, C. A. (2013). On the Import of Constraints in Complex Dynamical Systems. *Foundations of Science, 18*(4), 757–780.

Hooker, C. A., Penfold, H. B., & Evans, R. J. (1992). Towards a Theory of Cognition Under a New Control Paradigm. *Topoi, 11*(1), 71–88.

Hotchkiss, R. S., Strasser, A., McDunn, J. E., & Swanson, P. E. (2009). Cell Death in Disease: Mechanisms and Emerging Therapeutic Concepts. *New England Journal of Medicine, 361*(16), 1570–1583.

Huang, J. Y., & Bargh, J. A. (2014). The Selfish Goal: Autonomously Operating Motivational Structures as the Proximate Cause of Human Judgment and Behavior. *Behavioral and Brain Sciences*, *38*(01), 121–135.

Hurley, S. (2003). Animal Action in the Space of Reasons. *Mind and Language*, *18*(3), 231–257.

Illari, P. M., & Williamson, J. (2011). Mechanisms Are Real and Local. In P. M. Illiari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences* (pp. 818–844). Oxford: Oxford University Press.

Irwin, F. W. (1971). *Intentional Behavior and Motivation: A Cognitive Theory.* Philadelphia: Lippincott.

Janis, I. L., & Mann, L. (1977). *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*. New York: Free Press.

Jennings, N. R. (1993). Commitments and Conventions: The Foundation of Coordination in Multi-Agent Systems. *The Knowledge Engineering Review, 8*(3), 223–250.

Jones, D. M. (2017). *The Biological Foundations of Action*. London: Routledge.

Juarrero, A. (1999). *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge, MA: MIT Press.

Kálmán, R. E., & Bucy, R. S. (1961). New Results in Linear Filtering and Prediction Theory. *Journal of Basic Engineering, 83*(1), 95–108.

Kauffman, S. A. (1971). Articulation of Parts Explanation in Biology and the Rational Search for Them. In R. C. Buck & R. S. Cohen (Eds.), *PSA 1970* (pp. 257–272). Dordrecht: D. Reidel.

Keijzer, F. (2013). The *Sphex* Story: How the Cognitive Sciences Kept Repeating an Old and Questionable Anecdote. *Philosophical Psychology*, *26*(4), 502–519.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*, Cambridge, MA: Bradford.

Kenrick, D. T., Griskevicius, V., Neuberg, S. L., & Schaller, M. (2010). Renovating the Pyramid of Needs: Contemporary Extensions Built Upon Ancient Foundations. *Perspectives on Psychological Science, 5*(3), 292–314.

Khoo, M. C. K. (2018). *Physiological Control Systems: Analysis, Simulation, and Estimation* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Klee, R. L. (1984). Micro-Determinism and Concepts of Emergence. *Philosophy of Science, 51*(1), 44–63.

Klein, H. J., Becker, T. E., & Meyer, J. P. (Eds.). (2009). *Commitment in Organizations: Accumulated Wisdom and New Directions*. New York: Routledge.

Klein, H. J., Cooper, J. T., & Monahan, C. A. (2013). Goal Commitment. In E. A. Locke & G. P. Latham (Eds.), *New Developments in Goal Setting and Task Performance* (pp. 65–89). New York: Routledge.

Klein, H. J., Molloy, J. C., & Cooper, J. T. (2009). Conceptual Foundations: Construct Definitions and Theoretical Representations of Workplace Commitments. In H. J. Klein, T. E. Becker, & J. P. Meyer (Eds.), *Commitment in Organizations: Accumulated Wisdom and New Directions* (pp. 3–36). New York: Routledge.

Klein, H. J., & Park, H. M. (2016). Commitment as a Unidimensional Construct. In J. P. Meyer (Ed.), *Handbook of Employee Commitment* (pp. 15–27). Cheltenham, UK: Elgar.

Klinger, E. (1977). *Meaning and Void: Inner Experience and the Incentives in People's Lives*. Minneapolis: University of Minnesota Press.

Klinger, E. (1987). Current Concerns and Disengagement from Incentives. In F. Halisch & J. Kuhl (Eds.), *Motivation, Intention, and Volition* (pp. 337–347). Berlin: Springer-Verlag.

Klinger, E. (2013). Goal Commitments and the Content of Thoughts and Dreams: Basic Principles. *Frontiers in Psychology*, *4*.

Klinger, E., & Cox, W. M. (2004). Motivation and the Theory of Current Concerns. In W. M. Cox & E. Klinger (Eds.), *Handbook of Motivational Counseling: Concepts, Approaches, and Assessment* (pp. 3–27). Chichester, UK: Wiley.

Köhler, W. (1967). Some Gestalt Problems. In W. D. Ellis (Ed., Trans.), *A Source Book of Gestalt Psychology* (pp. 55–70). New York: Humanities. (Original work published 1922)

Krichmar, J. L. (2012). Design Principles for Biologically Inspired Cognitive Robotics. *Biologically Inspired Cognitive Architectures*, *1*, 73–81.

Kugler, P. N., Kelso, J. A. S., & Turvey, M. T. (1980). On the Concept of Coordinative Structures as Dissipative Structures: I. Theoretical Lines of Convergence. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in Motor Behavior* (pp. 3–47). Amsterdam: North-Holland.

Kuhlmann, M., & Glennan, S. (2014). On the Relation Between Quantum Mechanical and Neo-Mechanistic Ontologies and Explanatory Strategies. *European Journal for Philosophy of Science, 4*(2), 337–359.

Kuhn, W. (2009). A Functional Ontology of Observation and Measurement. In K. Janowicz, M. Raubal, & S. Levashkin (Eds.), *GeoSpatial Semantics* (pp. 26–43). Berlin: Springer-Verlag.

Lea, S. E. G. (1984). *Instinct, Environment, and Behaviour.* London: Methuen.

Lehrer, K. (1989). Metamental Ascent: Beyond Belief and Desire. *Proceedings and Addresses of the American Philosophical Association, 63*(3), 19–30.

Lepora, N. F., & Pezzulo, G. (2015). Embodied Choice: How Action Influences Perceptual Decision Making. *PLoS Computational Biology, 11*(4), e1004110.

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford: Oxford University Press.

Locke, E. A., & Latham, G. P. (1990). *A Theory of Goal Setting and Task Performance.* Englewood Cliffs, NJ: Prentice Hall.

Lydon, J. (1996). Toward a Theory of Commitment. In C. Seligman, J. M. Olson, & M. P. Zanna (Eds.), *The Psychology of Values* (pp. 191–214). Mahway, NJ: Erlbaum.

Lyon, P. (2006). The Biogenic Approach to Cognition. *Cognitive Processing*, *7*(1), 11–29.

MacCorquodale, K., & Meehl, P. E. (1954). Edward C. Tolman. In W. K. Estes, S. Koch, K. MacCorquodale, P. E. Meehl, C. G. Mueller, Jr., W. N. Schoenfeld & W. S. Verplanck, *Modern Learning Theory: A Critical Analysis of Five Examples* (pp. 177–266). New York: Appleton-Century-Crofts.

Machamer, P. (2004). Activities and Causation: The Metaphysics and Epistemology of Mechanisms. *International Studies in the Philosophy of Science, 18*(1), 27–39.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science, 67*(1), 1–25.

MacKay, D. M. (1952). Mentality in Machines III. *Proceedings of the Aristotelian Society, Supplementary Volumes, 26*(1), 61–86.

MacKay, D. M. (1962). The Use of Behavioural Language to Refer to Mechanical Processes. *British Journal for the Philosophy of Science, 13*(50), 89–103.

MacKay, D. M. (1964). Cybernetics. In J. Brierley (Ed.), *Science in its Context* (pp. 305–318). London: Heinemann.

MacKay, D. M. (1969). *Information, Mechanism, and Meaning*. Cambridge, MA: MIT Press.

Mackenroth, U. (2004). Robust Control Systems: Theory and Case Studies. Berlin: Springer-Verlag.

Maley, C. J., & Piccinini, G. (2017). A Unified Mechanistic Account of Teleological Functions for Psychology and Neuroscience. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (pp. 236–256). Oxford: Oxford University Press.

Marmodoro, A. (2017). Power Mereology: Structural versus Substantial Powers. In M. Paolini Paoletti & F. Orilia (Eds.), *Philosophical and Scientific Perspectives on Downward Causation* (pp. 110–127). New York: Routledge.

Mars, R. B., Sallet, J., Rushworth, M. F. S., & Yeung, N. (Eds.). (2011). *Neural Basis of Motivational and Cognitive Control*. Cambridge, MA: MIT Press.

Martin, C. B., & Pfeifer, K. (1986). Intentionality and the Non-Psychological. *Philosophy and Phenomenological Research, 46*(4), 531–554.

Massimi, M. (2018a). Four Kinds of Perspectival Truth. *Philosophy and Phenomenological Research, 96*(2), 342–359.

Massimi, M. (2018b). Perspectivism. In J. Saatsi (Ed.), *The Routledge Handbook of Scientific Realism* (pp. 164–175). New York: Routledge.

Maudlin, T. (2007). *The Metaphysics Within Physics.* Oxford: Oxford University Press.

McCall, S. (1987). Decision. *Canadian Journal of Philosophy, 17*(2), 261–287.

McClennen, E. F. (1990). *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.

McDowell, J. (1994). *Mind and World.* Cambridge, MA: Harvard University Press.

McFarland, D. J. (1989). *Problems of Animal Behaviour*. New York: John Wiley & Sons.

McFarland, D. J., & Bösser, T. (1993). *Intelligent Behavior in Animals and Robots*. Cambridge, MA: Bradford.

McFarland, D. J., & Sibly, R. M. (1975). The Behavioural Final Common Path. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 270*(907), 265–293.

McFarland, K., & Kalivas, P. W. (2003). Motivational Systems. In M. Gallagher & R. J. Nelson (Eds.), *Handbook of Psychology, Vol. 3: Biological Psychology* (pp. 379–403). Hoboken, NJ: John Wiley & Sons.

Metzinger, T. (2017). The Problem of Mental Action: Predictive Control without Sensory Sheets. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

Meyer, J. P. (Ed.). (2016). *Handbook of Employee Commitment*. Cheltenham, UK: Elgar.

Michael, J. A., Sebanz, N., & Knoblich, G. (2016). The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, *6*.

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202.

Millikan, R. G. (1989). In Defense of Proper Functions. *Philosophy of Science, 56*(2), 288–302.

Millikan, R. G. (1995). Pushmi-Pullyu Representations. *Philosophical Perspectives, 9*, 185–200.

Milsum, J. H. (1966). *Biological Control Systems Analysis*. New York: McGraw-Hill.

Minsky, M. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* New York: Simon & Schuster.

Mitchell, S. (2009). *Unsimple Truths: Science, Complexity and Policy.* Chicago: University of Chicago Press.

Mitchell, S. (2012). Emergence: Logical, Functional, and Dynamical. *Synthese, 185*, 171–186.

Molnar, G. (2003). *Powers: A Study in Metaphysics.* Oxford: Oxford University Press.

Montague, W. P. (1909). The True, the Good, and the Beautiful from a Pragmatic Standpoint. *Journal of Philosophy, Psychology and Scientific Methods, 6*(9), 233–238.

Montgomery, H. (1998). Decision Making and Action: The Search for a Dominance Structure. In M. Kofta, G. Weary, & G. Sedek (Eds.), *Personal Control in Action* (pp. 279–298). New York: Plenum.

Moreno, A., & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry.* Dordrecht: Springer.

Morgan, C. L. (1927). *Emergent Evolution* (2nd ed.). London: Williams and Norgate.

Morisano, D. (2013). Goal Setting in the Academic Arena. In E. A. Locke & G. P. Latham (Eds.), *New Developments in Goal Setting and Task Performance* (pp. 495–506). New York: Routledge.

Morris, C. (1946). *Signs, Language, and Behavior*. New York: Prentice-Hall.

Mossio, M., Saborido, C., & Moreno, A. (2009). An Organizational Account of Biological Functions. *British Journal of Philosophy of Science, 60*(4), 813–841.

Mugg, J. (2016). The Dual-Process Turn: How Recent Defenses of Dual-Process Theories of Reasoning Fail. *Philosophical Psychology*, *29*(2), 300–309.

Mugg, J. (2018). The Sound-Board Account of Reasoning: A One-System Alternative to Dual-Process Theory. *Philosophical Psychology*, *31*(7), 1046–1073.

Mumford, S. (2004). *Laws in Nature.* London: Routledge.

Mumford, S., & Anjum, R. L. (2011). *Getting Causes from Powers.* Oxford: Oxford University Press.

Murphy, N., & Brown, W. S. (2007). *Did My Neurons Make Me Do It? Philosophical and Neurobiological Perspectives on Moral Responsibility and Free Will*. Oxford: Oxford University Press.

Murphy, R. R. (2000). *Introduction to AI Robotics*. Cambridge, MA: Bradford.

Muse, D., & Wermter, S. (2009). Actor-Critic Learning for Platform-Independent Robot Navigation. *Cognitive Computation*, *1*(3), 203–220.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation.* New York: Harcourt, Brace & World.

Nagel, E. (1977). Functional Explanations in Biology. *Journal of Philosophy, 74*(5), 280–301.

Nagel, T. (1970). *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

Nagel, T. (1986). *The View from Nowhere*. Oxford: Oxford University Press.

Nanay, B. (2010). A Modal Theory of Function. *Journal of Philosophy, 107*(8), 412–431.

Neander, K. (1991). Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of Science, 58*, 168–184.

Neander, K. (1995). Misrepresenting & Malfunctioning. *Philosophical Studies, 79*, 109–141.

Neander, K., & Rosenberg, A. (2012). Solving the Circularity Problem for Functions: A Response to Nanay. *Journal of Philosophy, 109*(10), 613–622.

Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology.* San Francisco: W. H. Freeman.

Nicolis, G., & Prigogine, I. R. (1977). *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order Through Fluctuations.* New York: Wiley.

Nogueira, E., Fidalgo, M., Molnar, A., Kyriakis, J., Force, T., Zalvide, J., & Pombo, C. M. (2008). SOK1 Translocates from the Golgi to the Nucleus upon Chemical Anoxia and Induces Apoptotic Cell Death. *Journal of Biological Chemistry, 283*(23), 16248–16258.

Nyberg, I. (2009). Can Moral Norms Be Derived from Nature? The Incompatibility of Natural Scientific Investigation and Moral Norm Generation. In M. J. Cherry (Ed.), *The Normativity of the Natural: Human Goods, Human Virtues, and Human Flourishing* (pp. 175–196). Dordrecht: Springer.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology, 68*, 73–100.

O'Shaughnessy, B. (1980). *The Will: A Dual Aspect Theory*. Cambridge: Cambridge University Press.

Okamoto, K., Kondo-Okamoto, N., & Ohsumi, Y. (2009). Mitochondria-Anchored Receptor Atg32 Mediates Degradation of Mitochondria via Selective Autophagy. *Developmental Cell, 17*, 87–97.

Okasha, S. (2009). Causation in Biology. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford Handbook of Causation* (pp. 707–725). Oxford: Oxford University Press.

Pacherie, E. (2008). The Phenomenology of Action: A Conceptual Framework. *Cognition*, *107*(1), 179–217.

Palmer, G. B. (2007). Cognitive Linguistics and Anthropological Linguistics. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 1045–1073). Oxford: Oxford University Press.

Parfit, D. (1997). Reasons and Motivation I. *Aristotelian Society Supplementary Volume*, *71*(1), 99–130.

Pattee, H. H. (1970). The Problem of Biological Hierarchy. In C. H. Waddington (Ed.), *Towards a Theoretical Biology 3: Drafts* (pp. 117–136). Edinburgh: Edinburgh University Press.

Pattee, H. H. (1971). Physical Theories of Biological Co-ordination. *Quarterly Reviews of Biophysics, 4*(2 & 3), 255–276).

Pattee, H. H. (1972). Laws and Constraints, Symbols and Languages. Reprinted in H. H. Pattee & J. Rączaszek-Leonardi (Eds.), *Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary* (pp. 81–89). Dordrecht: Springer, 2012.

Pattee, H. H. (1973a). The Physical Basis and Origin of Hierarchical Control. Reprinted in H. H. Pattee & J. Rączaszek-Leonardi (Eds.), *Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary* (pp. 91–110). Dordrecht: Springer, 2012.

Pattee, H. H. (1973b). Physical Problems of Decision-Making Constraints. Reprinted in H. H. Pattee & J. Rączaszek-Leonardi (Eds.), *Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary* (pp. 69–79). Dordrecht: Springer, 2012.

Pattee, H. H. (1973c). Physical Problems of the Origin of Natural Controls. In A. Locker (Ed.), *Biogenesis, Evolution, Homeostasis* (pp. 41–49). Berlin: Springer-Verlag.

Pattee, H. H. (1977). Dynamic and Linguistic Modes of Complex Systems. *International Journal of General Systems, 3*, 259–266.

Pattee, H. H. (1982). Cell Psychology: An Evolutionary Approach to the Symbol-Matter Problem. Reprinted in H. H. Pattee & J. Rączaszek-Leonardi (Eds.), *Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary* (pp. 165–179). Dordrecht: Springer, 2012.

Pattee, H. H. (1991). Measurement-Control Heterarchical Networks in Living Systems. *International Journal of General Systems, 18*(3), 213–221.

Pattee, H. H. (1992). The Measurement Problem in Physics, Computation, and Brain Theories. In M. E. Carvallo (Ed.), *Nature, Cognition and System II* (pp. 179–192). Dordrecht: Kluwer.

Pattee, H. H. (1996). The Problem of Observables in Models of Biological Organizations. Reprinted in H. H. Pattee & J. Rączaszek-Leonardi (Eds.), *Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary* (pp. 245–259). Dordrecht: Springer, 2012.

Pepper, S. C. (1958). *The Sources of Value*. Berkeley: University of California Press.

Peter, F., & Schmid, H. B. (Eds.). (2007). *Rationality and Commitment*. Oxford: Oxford University Press.

Pezzulo, G. (2012). An Active Inference View of Cognitive Control. *Frontiers in Psychology*, *3*(478).

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical Active Inference: A Theory of Motivated Control. *Trends in Cognitive Sciences*, *22*(4), 294–306.

Piccinini, G. (2015). *Physical Computation*. Oxford: Oxford University Press.

Pitkin, W. B. (1912). Some Realistic Implications of Biology. In E. B. Holt, W. T. Marvin, W. P. Montague, R. B. Perry, W. B. Pitkin, & E. G. Spaulding, *The New Realism: Cooperative Studies in Philosophy* (pp. 377–467). New York: Macmillan.

Polányi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy.* New York: Harper Torchbooks.

Pollock, J. L. (2006). *Thinking about Acting: Logical Foundations for Rational Decision Making*. Oxford: Oxford University Press.

Polman, E., & Russo, J. E. (2012). Commitment to a Developing Preference and Predecisional Distortion of Information. *Organizational Behavior and Human Decision Processes*, *119*(1), 78–88.

Popper, K. (2002). *Conjectures and Refutations: The Growth of Scientific Knowledge* (6th ed.). London and New York: Routledge Classics.

Portmore, D. W. (2013). Agent-Relative vs. Agent-Neutral. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics* (pp. 162–171). Malden, MA: Wiley-Blackwell. https://doi.org/10.1002/9781444367072.wbiee043

Powers, W. T. (1973). *Behavior: The Control of Perception*. New York: Aldine.

Premack, D. (2007). Human and Animal Cognition: Continuity and Discontinuity. *Proceedings of the National Academy of Sciences*, *104*(35), 13861–13867.

Preston, B. (1998). Why is a Wing Like a Spoon? A Pluralist Theory of Function. *Journal of Philosophy, 95*(5), 215–254.

Price, H. H. (1969). *Belief*. London: George Allen & Unwin.

Purich, D. L. (2010). *Enzyme Kinetics: Catalysis and Control*. Amsterdam: Academic Press.

Railton, P. (2006). Normative Guidance. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 1, pp. 3–33). Oxford: Clarendon.

Rapport, N., & Overing, J. (2000). *Social and Cultural Anthropology: The Key Concepts*. London: Routledge.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The Basal Ganglia: A Vertebrate Solution to the Selection Problem? *Neuroscience, 89*(4), 1009–1023.

Redshaw, J., & Bulley, A. (2018). Future-Thinking in Animals: Capacities and Limits. In G. Oettingen, A. T. Sevincer, & P. M. Gollwitzer (Eds.), *The Psychology of Thinking about the Future* (pp. 31–51). New York: Guilford.

Reeve, J. (2015). *Understanding Motivation and Emotion* (6th ed.). Hoboken, NJ: Wiley.

Rescher, N. (1969). The Concept of Control. In his *Essays in Philosophical Analysis* (pp. 327–353). Pittsburgh: University of Pittsburgh Press.

Rescher, N. (1970). On the Characterization of Actions. In M. Brand (Ed.), *The Nature of Human Action* (pp. 247–254). Glenview, IL: Scott, Foresman and Company.

Robertson, R. J., & Powers, W. T. (1990). *Introduction to Modern Psychology: The Control-Theory View*. Gravel Switch, KY: CSG.

Rosin, P. L., & Rana, O. F. (2004). Agent-Based Computer Vision. *Pattern Recognition, 37*, 627–629.

Ross, D. (2000). Rainforest Realism: A Dennettian Theory of Existence. In D. Ross, A. Brook, & D. Thompson (Eds.), *Dennett's Philosophy: A Comprehensive Assessment* (pp. 147–168). Cambridge, MA: Bradford.

Ross, D. (2018). Game Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 Ed.). https://plato.stanford.edu/archives/fall2018/entries/game-theory/

Ross, D., Ladyman, J., & Collier, J. (2007). Rainforest Realism and the Unity of Science. In J. Ladyman & D. Ross, *Every Thing Must Go: Metaphysics Naturalized* (pp. 190–257). Oxford: Oxford University Press.

Ross, D., Ladyman, J., & Spurrett, D. (2007). Causation in a Structural World. In J. Ladyman & D. Ross, *Every Thing Must Go: Metaphysics Naturalized* (pp. 258–297). Oxford: Oxford University Press.

Ross, L. N. (forthcoming-a). Causal Control: A Rationale for Causal Selection. In C. K. Waters & J. Woodward (Eds.), *Minnesota Studies in the Philosophy of Science, Vol. XXI: Philosophical Perspectives on Causal Reasoning in Biology*. Minneapolis: University of Minnesota Press.

Ross, L. N. (forthcoming-b). Causal Concepts in Biology: How Pathways Differ from Mechanisms and Why it Matters. *British Journal for the Philosophy of Science*. http://philsci-archive.pitt.edu/14432/

Rosu, V., & Hughes, K. T. (2006). σ^{28}-Dependent Transcription in *Salmonella enterica* is Independent of Flagellar Shearing. *Journal of Bacteriology, 188*(14), 5196–5203.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Ryle, G. (1949). *The Concept of Mind* (60th Anniversary Ed.). London: Routledge. (Reprinted 2009)

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton, NJ: Princeton University Press.

Salthe, S. N. (2007). The Natural Philosophy of Work. *Entropy, 9*, 83–99.

Sayre, K. (1976). *Cybernetics and the Philosophy of Mind*. Atlantic Highlands, NJ: Humanities.

Schapiro, T. (2009). The Nature of Inclination. *Ethics*, *119*(2), 229–256.

Schlosser, M. E. (2007). Basic Deviance Reconsidered. *Analysis, 67*(3), 186–194.

Schroeder, T. (2004a). Functions from Regulation. *The Monist, 87*(1), 115–135.

Schroeder, T. (2004b). *Three Faces of Desire*. Oxford: Oxford University Press.

Schultheiss, O. C., & Brunstein, J. C. (2010). (Eds.). *Implicit Motives*. New York: Oxford.

Schultheiss, O. C., Strasser, A., Rösch, A. G., Kordik, A., & Graham, S. C. C. (2012). Motivation. In V. S. Ramachandran (Ed.), *Encyclopedia of Human Behavior* (2nd Ed., pp. 650–656). London: Academic.

Searle, J. R. (1979). *Expression and Meaning*. Cambridge: Cambridge University Press.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Selfridge, O. G., & Neisser, U. (1960). Pattern Recognition by Machine. *Scientific American, 203*(2), 60–68.

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating into the Future or Driven by the Past. *Perspectives on Psychological Science, 8*(2), 119–141.

Sellars, W. (1954). Some Reflections on Language Games. Reprinted in his *Science, Perception, and Reality* (pp. 321–358). Atascadero, CA: Ridgeview, 1963.

Sen, A. K. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, *6*(4), 317–344.

Sen, A. K. (2005). Why Exactly Is Commitment Important for Rationality? *Economics and Philosophy*, *21*(1), 5–14.

Shadlen, M. N., & Kiani, R. (2013). Decision Making as a Window on Cognition. *Neuron*, *80*(3), 791–806.

Shah, Y. T. (2018). *Thermal Energy: Sources, Recovery, and Applications*. Boca Raton, FL: CRC.

Sharot, T. (2012). Predicting Emotional Reactions: Mechanisms, Bias and Choice. In *Neuroscience of Preference and Choice* (pp. 53–72). Amsterdam: Academic Press.

Shepherd, J. (2014). The Contours of Control. *Philosophical Studies, 170*, 395–411.

Shoemaker, D. W. (2003). Caring, Identification, and Agency. *Ethics*, *114*(1), 88–118.

Shoemaker, S. (1998). Causal and Physical Necessity. *Pacific Philosophical Quarterly, 79*, 59–77.

Shpall, S. (2016). The Calendar Paradox. *Philosophical Studies*, *173*(3), 801–825.

Simon, H. A. (1979). *Models of Thought*. New Haven, CT: Yale University Press.

Sinnott, E. W. (1961). *Cell and Psyche: The Biology of Purpose*. New York: Harper & Row.

Sloman, A. (1993). The Mind as a Control System. *Royal Institute of Philosophy Supplements, 34*, 69–110.

Sloman, A. (1996). Towards a General Theory of Representations. In D. Peterson (Ed.), *Forms of Representation: An Interdisciplinary Theme for Cognitive Science* (pp. 118–140). Exeter, UK: Intellect.

Smart, J. J. C. (1963). *Philosophy and Scientific Realism.* London: Routledge and Kegan Paul.

Sober, E. (1984). *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. Cambridge, MA: Bradford.

Stalnaker, R. C. (1984). *Inquiry*. Cambridge, MA: Bradford.

Stanovich, K. (1999). *Who is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Staude, M. (1986). Wanting, Desiring, and Valuing: The Case against Conativism. In J. Marks (Ed.), *The Ways of Desire* (175–195). Chicago, IL: Precedent.

Stear, E. B. (1987). Control Paradigms and Self-Organization in Living Systems. In F. E. Yates, A. Garfinkel, D. O. Walter, & G. B. Yates (Eds.), *Self-Organizing Systems: The Emergence of Order* (pp. 351–397). New York: Plenum.

Stein, D. L. (Ed.). (1989). *Lectures in the Sciences of Complexity.* Redwood City, CA: Addison-Wesley.

Stephens, C. (1998). Bacterial Sporulation: A Question of Commitment? *Current Biology*, *8*(2), R45–R48.

Sterelny, K. (1995). Basic Minds. *Philosophical Perspectives, 9*, 251–270.

Sterelny, K. (2001). The Evolution of Agency. In his *The Evolution of Agency and Other Essays* (pp. 260–287). Cambridge: Cambridge University Press.

Sterelny, K. (2003). *Thought in a Hostile World: The Evolution of Human Cognition*. Malden, MA: Blackwell.

Steward, H. (2009). Animal Agency. *Inquiry, 52*(3), 217–231.

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford: Clarendon.

Stich, S. P. (1978). Do Animals Have Beliefs? *Australian Journal of Philosophy, 57*,15–28.

Strawson, P. F. (1959). *Individuals: An Essay in Descriptive Metaphysics.* London: Routledge.

Sullivan-Bissett, E. (2017). Biological Function and Epistemic Normativity. *Philosophical Explorations, 20*(sup1), 94–110.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. Cambridge, MA: Bradford.

Szpunar, K. K., Shrikanth, S., & Schacter, D. L. (2018). Varieties of Future-Thinking. In G. Oettingen, A. T. Sevincer, & P. M. Gollwitzer (Eds.), *The Psychology of Thinking about the Future* (pp. 52–67). New York: Guilford.

Tal, E. (2015). Measurement in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 Ed.). https://plato.stanford.edu/archives/fall2017/entries /measurement-science/

Taylor, J. R. (2005). *Classical Mechanics.* Sausalito, CA: University Science Books.

Taylor, R. (1966). *Action and Purpose*. Englewood Cliffs, NJ: Prentice-Hall.

Taylor, S. E., & Pham, L. B. (1996). Mental Simulation, Motivation, and Action. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The Psychology of Action: Linking Cognition and Motivation to Behavior* (pp. 219–235). New York: Guilford.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Belknap.

Toates, F. M. (1975). *Control Theory in Biology and Experimental Psychology*. London: Hutchinson.

Toates, F. M. (1986). *Motivational Systems*. Cambridge: Cambridge University Press.

Tolman, E. C. (1936). Connectionism; Wants, Interests, and Attitudes. *Journal of Personality*, *4*(3), 245–253.

Toribio, J. (2007). Review of *Mind and Supermind* by Keith Frankish. *Philosophical Quarterly, 57*(226), 139–142.

Turner, S. P. (1994). *The Social Theory of Practices: Tradition, Tacit Knowledge and Presuppositions*. Cambridge: Polity.

Tweedale, J., Ichalkaranje, N., Sioutis, C., Jarvis, B., Consoli, A., & Phillips-Wren, G. (2007). Innovations in Multi-Agent Systems. *Journal of Network and Computer Applications*, *30*(3), 1089–1115.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Clarendon.

Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon.

Varela, F. J. (1997). Patterns of Life: Intertwining Identity and Cognition. *Brain and Cognition, 34*, 72–87.

Vitz, R. (2008). Doxastic Voluntarism. In *The Internet Encyclopedia of Philosophy*, https://www.iep.utm.edu/doxa-vol/ (Retrieved on January 1, 2019).

Walsh, D. M. (1996). Fitness and Function. *British Journal for the Philosophy of Science, 47*, 553–574.

Walsh, D. M. (2015). *Organisms, Agency, and Evolution*. Cambridge: Cambridge University Press.

Wason, P. C., & Evans, J. St. B. T. (1974). Dual Process in Reasoning? *Cognition, 3*(2), 141–154.

Watson, G. (1975). Free Agency. *Journal of Philosophy*, *72*(8), 205–220.

Weirich, P. (2013). Preference. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics* (pp. 4041–4051). Malden, MA: Wiley-Blackwell. https://doi.org/10.1002 /9781444367072.wbiee744

Weiss, P. (1938). *Reality.* Princeton, NJ: Princeton University Press.Weissman, D. (1965). *Dispositional Properties.* Carbondale: Southern Illinois University Press.

Wimsatt, W. C. (1972). Teleology and the Logical Structure of Function Statements. *Studies in the History and Philosophy of Science, 3*(1), 1–80.

Wimsatt, W. C. (1976). Reductive Explanation: A Functional Account. *Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1974*, 671–710.

Wimsatt, W. C. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy, Suppl. Vol. 20*, 207–274.

Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.* Cambridge, MA: Harvard University Press.

Winning, J. (forthcoming). Mechanistic Causation and Constraints: Perspectival Parts and Powers, Non-Perspectival Modal Patterns. *British Journal for the Philosophy of Science.* doi: 10.1093/bjps/axy042

Winning, J., & Bechtel, W. (2018). Rethinking Causality in Biological and Neural Mechanisms: Constraints and Control. *Minds and Machines, 28*(2), 287–310.

Winning, J., & Bechtel, W. (forthcoming). Being Emergence vs. Pattern Emergence: Complexity, Control, and Goal-Directedness in Biological Systems. In S. C. Gibb, R. F. Hendry, & T. Lancaster (Eds.), *The Routledge Handbook of Emergence*. London: Routledge.

Witherington, D. C. (2011). Taking Emergence Seriously: The Centrality of Circular Causality for Dynamic Systems Approaches to Development. *Human Development*, *54*(2), 66–92.

Wolf, S. (1990). *Freedom within Reason*. New York: Oxford University Press.

Wolpert, D. M., Ghahramani, Z, & Jordan, M. I. (1995). An Internal Model for Sensorimotor Integration. *Science, 269*(5232), 1880–1882.

Woodger, J. H. (1929). *Biological Principles*. London: Routledge & Kegan Paul.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenal Research, 77*(1), 193–212.

Wooldridge, M. J. (2009). *An Introduction to MultiAgent Systems* (2nd Ed.). Chichester, UK: Wiley.

Wooldridge, M. J., & Jennings, N. R. (1995). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review, 10*(2), 115–152.

Ybarra, O., & Winkielman, P. (2012). On-Line Social Interactions and Executive Functions. *Frontiers in Human Neuroscience*, *6*.

Yolum, P., & Singh, M. P. (2002). Commitment Machines. In J.-J.Ch. Meyer & M. Tambe (Eds.), *Intelligent Agents VIII: Agent Theories, Architectures, and Languages* (pp. 235–247). Berlin: Springer-Verlag.

Youle, R. J., & Narendra, D. P. (2011). Mechanisms of Mitophagy. *Nature Reviews Molecular Cell Biology, 12*, 9–14.