INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®
RICE UNIVERSITY

A Theory of Willpower

by

Robert B. Tierney

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE

Doctor of Philosophy

APPROVED THESIS COMMITTEE:

Baruch Brody, Professor of Philosophy

Richard Grandy, Carolyn and Fred McManis
Professor of Philosophy

David Schneider, Professor of Psychology

HOUSTON, TEXAS

DECEMBER, 2001
December, 2001

ABSTRACT

A Theory of Willpower

by

Robert B. Tierney

This dissertation explores the phenomenology and folk theory of willpower, critically assesses existing scientific theories of willpower that seek to explain the phenomenology and folk theory, and proposes a new scientifically useful theory of willpower. Earlier theories reliant upon such mechanisms or techniques as control of attention, rulemaking and private side bets are shown to be inadequate to explain crucial features of certain behavioral performance deemed to be the exhibition of willpower. This dissertation sets out to a new theory capable of explaining these key features of willpower in terms of a new model of private side bets involving the staking of status ascriptions of strength of will upon the success of the relevant behavioral performances.
Table of Contents

Chapter One: The Project

I. General Introduction ................................................................. 1

II. Why Willpower? ........................................................................... 1
   A. Introduction ............................................................................. 1
   B. Theoretical and Practical Reasons for an Interest in Willpower .... 2
   C. The Changing Fortunes of Willpower in Explanations of Behavior .... 10

III. The Explanatory Project ............................................................... 19
   A. The Basic Strategy ................................................................. 19
   B. Basic Motivational Assumptions ............................................. 27

IV. Conclusion .................................................................................. 35

Chapter Two: Folk and Scientific Theories of Willpower

I. Chapter Introduction ................................................................. 36

II. James on the Folk Theory of Willpower .................................... 36
   A. The Phenomenology of Willpower ......................................... 36
   B. The Folk Theory of Willpower .............................................. 38
   C. Initial Criticisms of James’ Account of the Phenomenology and Folk Theory of Willpower ............................................. 39

III. James’ Scientific Theories of Willpower .................................... 41
   A. Introduction ............................................................................. 41
   B. James’ Primary Theory of Willpower ..................................... 42
   C. Problems with James’ Primary Theory of Willpower ................. 45
   D. James’ Supplementary Theory of Willpower ............................ 51
   E. Problems with James’ Supplementary Theory of Willpower ......... 52
F. The Feeling of Effort .................................................................56
G. The Effect of James' Scientific Theories on His Description of the Folk Concept of Willpower ......................................................61

IV. Two Concepts of Willpower .....................................................63
V. Theories of Input Willpower .....................................................66
VI. Conclusions Regarding James' Theories of Willpower ..................71
VII. Summing Up The Concept of Output Willpower .......................73
VIII. Summing Up the Chapter and Moving Forward .......................87

Chapter Three: Personal Rules and Private Side Bets

I. Introduction .............................................................................89
II. The Basic Personal Rules Mechanism ......................................89
III. Some Varieties of Personal Rules Mechanisms .......................96
   A. Overview ...........................................................................96
   B. Problems With Intentionally Employed Predictive Personal Rules ....97
   C. Prescriptive Personal Rules .................................................102
      1. Overview ....................................................................102
      2. The Impossibility of Prescriptive Personal Rules .................103

IV. Further Variations on a Theme ..............................................119
   A. Introduction .....................................................................119
   B. Nozick's Version of Personal Rules ...................................119
   C. Bratman's Version of Personal Rules ...................................128
   D. A Final Variation: Will-Based Side Bets .............................132

V. Conclusion ............................................................................136
Chapter Four
A New Theory of Output Willpower

I. Introduction ................................................................. 139

II. "Betting" as a Performative ............................................. 140

III. Examples of Staking One's Conventional Status From Honor Cultures .... 144
    A. Early and Mid-Twentieth Century Japan .......................... 144
    B. The World of the Norse Sagas ..................................... 145

IV. The Self-Enforcing Nature of Status Staking ............................ 147

V. Further Refinements of the Mechanism of Status Staking .................. 153
    A. Introduction .......................................................... 153
    B. Ways of Staking One's Status ..................................... 154
    C. Staking by Degrees .................................................. 160
    D. Conclusion ............................................................ 163

VI. The Status of the Strong-Willed ....................................... 163

VII. Functional Isomorphism .................................................. 167

VIII. The View From Inside ................................................... 169
    A. Introduction .......................................................... 169
    B. Accepting Status Staking on Its Own Terms ...................... 169
    C. The Metaphysical View of the Will .................................. 172

IX. Summation ................................................................. 182

X. General Objections, Conclusions and Prospects ........................... 182

XI. Conclusion ........................................................................ 187

Bibliography ........................................................................ 188
Appendix A

The Iterated Prisoner's Dilemma Analysis of Personal Rules ................................202
Chapter One

The Project

I. General Introduction

The aim of this dissertation is to explain willpower. More accurately, the aim is to provide an account of several features of a subset of cases of behavioral performance that could appropriately be called examples of the exercise of willpower, and to do so within a framework of explanation common to most of the behavioral sciences. Before beginning down the path of this explanation, it is worthwhile to first say a bit about what the target of this explanation is, why philosophers should care about explaining it, and what kind of explanation I plan to offer.

II. Why Willpower?

A. Introduction

Why should a philosopher be interested in explaining willpower? In a sense, the whole of this dissertation will be an effort to demonstrate that the subject is philosophically interesting by fleshing out just what the subject is and showing how it has figured in explanations of human behavior by ordinary persons, philosophers and psychologists. However, at this point, I can at least make a preliminary effort to clarify the target of the explanation and offer some general reasons why the topic should be of interest to philosophers and others interested in understanding human behavior. I will also set out the specific reasons as to why this topic interests this writer in hopes that these particular concerns might resonate with some readers. Further, I will provide a very brief summary of how willpower, along with the related topics of volition and, more generally, the exercise of the will, have figured in the history of philosophy and
psychology. I will then offer some evidence that these topics have gained a renewed interest among philosophers, psychologists and others concerned with explaining human action. This should at least show that philosophers and others seeking to explain the actions of person have historically been interested in the topic and continue to be interested in the topic today. More generally, this part of the work should also begin to shed light on why such persons were and are interested in willpower. Together, I hope that the various considerations set out in this section will show willpower to be a topic that is interesting and important for philosophers and others concerned with human action.

B. Theoretical and Practical Reasons for an Interest in Willpower

To begin, I will offer a few general reasons as to why willpower is a topic of philosophical interest. First, willpower is of interest to the extent that it figures in the explanation of human action. As we shall see later in this chapter, there is a long philosophical history of efforts to explain human action in terms of causally efficacious acts of willing. Further, if one is interested in explaining human action, and if, as I argue in what follows, willpower plays a significant role in human action, then it follows that one should also have an interest in willpower. A good part of the work of the next chapter will be to show how folk theories of behavior assign an important role to willpower in human behavior. The work of the remainder of the dissertation will be to see whether and how this concept can be explained in more scientifically acceptable terms.

The reason for proposing an account that can be woven into the fabric of the behavioral sciences is the hope that willpower can be taken out of the realm of the
mysterious and intractable and placed within the realm of intelligible and testable theories with both theoretical and practical benefits. The hope is that in so explaining willpower, we can prevent it being dismissed as mere myth by those who purport to explain human behavior when there is, in fact, something there that cries out for explanation. At the same time, finding a credible and comprehensible explanation for willpower may serve as a sort of prophylactic against persons simply adopting wholesale and without reflection the concept of willpower in its more romantic or magical versions, i.e. as a kind of magical force at the command of the individual operating outside the ordinary laws of nature. Finally, I hope to remove a source of philosophical puzzlement by bringing the explanation down to earth, so to speak, by explaining willpower in terms of a rather pedestrian psychological mechanism. What I mean by all this will become clearer in the course of this chapter. For now, I will simply state that this leads to the second kind of reason as to why a scientifically relevant account of willpower is important: an improved understanding of willpower may ultimately yield a payoff in terms of our practical efforts to lead richer, more intelligible and more effective lives.

With respect to these practical concerns, one set of issues in particular helps in significant measure to account for the recent increase in interest in willpower: concerns over irrational or self-destructive behaviors that are apparently connected with deficits in, or aberrations of, motivation. Perhaps the most important, or at least the most publicly recognized, of these behavioral problems is that of addiction in all its myriad forms. There seems to be an increasing concern among the public, those charged with treating or otherwise dealing with persons suffering from addictions, and the addicts themselves to understand and appropriately deal with addiction.
As shall be set forth in more detail in what follows, both ordinary folk theories and more sophisticated psychological theories have looked to willpower as a solution to problems concerning motivational deficits. Through the course of this dissertation, I will offer some evidence that I hope is at least suggestive of why our understandings and misunderstandings of willpower can in direct and concrete ways benefit or harm those suffering from repeated failures of will. However, the application of the theory of willpower that I will develop will not be a part of this dissertation, though I hope that I can make this a subject of future work.

For the moment, I will simply try to sum up by pointing out the application of the twin dangers described above in the context of motivational problems such as addiction. On the one hand, there are those who warn that explaining away the power of will as mere myth or as not causally efficacious runs the practical risk of fostering a sense of helplessness or fatalism that tends to generate problem behaviors including alcohol abuse.¹ On the other hand, those who embrace a magical or romantic conception of willpower such that they conceive of it as a force of more or less unlimited motivational power may also be courting disaster. Practicing alcoholics, for instance, are persons who, as a group, are often credited with being rather too willful in most domains of life and as having a rather inflated concept of willpower. It is a commonplace that before seeking treatment for their addiction, they repeatedly try, and repeatedly fail, to overcome their

alcoholism by resort to efforts of willpower. Among such persons, slogans such as the much-publicized injunction “Just say no!” become little more than a source of bitterly ironic humor. These considerations suggest that a more developed understanding of willpower may benefit those who must take practical action to address issues of public policy such as health care and criminal justice.

Third, a theory of agency is an important building block in a variety of other theories. These include theories in the domain of ethics, political philosophy, political science, sociology, anthropology, economics and law. To the extent that willpower is an important part of the explanation of intentional behavior and, nonetheless, is left out of such higher-order theories, these theories may go astray in assessing how populations, communities, firms, nations or states should conduct themselves or actually do conduct themselves.

But, of course, this third reason implies a fourth reason why the study of willpower is of interest to philosophers and social scientists. There is more than the mere assessment and recommendation of academics or experts at stake in these higher-order theories. The conclusions reached by experts in these disciplines may become embodied in social movements, organizational decisions, policies and laws. Our theories of group behavior may issue not only in assessments and recommendations, but also in binding mandates and programs of action.

---

In light of the third and fourth reasons I have given for an interest in the functioning of willpower, it is worth noting that in recent years Elster has criticized the general approach of much of the work done in the social sciences, arguing that we can best facilitate the understanding of human action by focusing on discrete behavioral mechanisms, as opposed to general theories of motivation and behavior. To the extent that a more general theory can be offered, it is only in light of the results obtained from these more focused analyses. Elster writes that the method that he deems to be "exactly right," at least for some purposes, is "to understand economics through social psychology, and social psychology through history." In this way, according to Elster, we can best move from the particulars of history which illuminate specific behavioral mechanisms which may be operative or manifested in a particular way only in a certain historical context, to a more abstract psychological conceptualization of such mechanisms, and only then, in light of such particular studies, attempt the more abstract

3 Elster has written so much that is of interest to economists, sociologists, philosophers, psychologists, historians, and political theorists, that I cannot claim to know how to categorize his work. A recent publication states that he holds a chair in "social science." Jon Elster, Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints, (Cambridge: Cambridge University Press, 2000).

4 Over time, Elster seems to have become increasingly skeptical about the possibility of a meaningful general theory of human behavior and has become more direct in his advocacy of a mechanism-centered approach. See Jon Elster, Ulysses and the Sirens: Studies in Rationality and Irrationality, (Cambridge: Cambridge University Press, 1979, revised 1984).


5 Elster, Political Psychology at 57.
generalizations of economics. I am largely in agreement with Elster in this regard, except that I would prefer to alter the description to read “history or anthropology” in Elster’s formulation, since anthropology also concerns itself with the particulars of human behavior in all its variety. Still, both disciplines perform analogous functions in that they are concerned with the particulars human action in specific times and places. I would also add that it doesn’t hurt to have a philosopher around to tidy things up and offer a new hypothesis from time to time, as I hope to show.

Ideally, this dissertation constitutes an example of such a study of a particular mechanism in the general vein that Elster advocates. However, I must approach the work from the vantage point of a philosopher, and am, therefore, somewhat more reliant than the historian and the anthropologist upon others to assemble the particular facts for me. Still, I do have the benefit that the mechanism targeted for explanation is operative at the present point in time in the society of which I am a member. So I don’t have to go far before encountering relevant “data.” Moreover, my account of willpower will be reliant upon both the analytical tools and the specific analytical results developed by philosophers. To put the point somewhat tendentiously, one might say that philosophers have already been doing the fieldwork on this particular topic. Further, as I shall argue below, philosophy has an important role to play in such investigations of human action even when such accounts are intended to be a part of the broader project of the human sciences.

In attempting to show why a philosopher should be interested in a “scientifically acceptable” explanation of willpower, I can also offer some account of how the topic of

---

6 Ibid. at Introduction, Ch. 1.
willpower fits within larger theoretical concerns in my own work—or, more accurately, how I hope it may eventually figure into such work—in accord with the move from the particular to the general, and the individual to the group, outlined by Elster. Briefly stated, I will argue in this dissertation that the exercise of willpower depends in very important ways upon a certain kind of convention status ascription. Specifically, I argue that the operative status is that of a powerful and efficacious agent who can successfully carry out actions by dint of willpower where he\(^7\) would otherwise fail if the matter were left to inclination alone.

This connects up with arguments in my master’s thesis. There I argued that the normative account of rationality sometimes known as “Self-Interest Theory” is such that, were an agent to embrace it and use it as a guide to action, it would alienate him from his own projects and commitments—the endeavors and obligations that give meaning to his life—in that he would necessarily have to regard them from the second-order perspective of Self-Interest Theory as merely means to maximizing his own satisfaction.\(^8\) Further, and more to the point, I argue that ultimately the agent would have to regard his very character instrumentally. This is where the connection with willpower comes in (or so I hope). For I argue that operating on the basis of Self-Interest Theory tends to make fungible the agent’s own character traits—any of them can, so to speak, be traded out when the balance of interests favors such a move. But the effective functioning of

---

\(^7\) For convenience I have chosen to avoid the cumbersome usage of always referring to both genders. Initially the dissertation was drafted using only feminine pronouns. However, this resulted in making the work unnecessarily complicated and left it sounding unnecessarily contrived in that nearly all of the passages from other writers cited herein that give examples of acts of willpower use masculine pronouns. Weaving those texts into this one was made easier by use of the masculine. These terms are to be understood as neutral except where the context indicates otherwise.

\(^8\) See, e.g. Robert B. Tierney, Seeking the Standpoint of Rationality, (Master’s thesis, Rice University, 1993).
willpower may depend upon its operative self-conception being a fixed (or at least not a fungible) aspect of character or self. If that is the case, then operating on the basis of Self-Interest Theory may undermine the efficacy of a powerful behavioral resource for enabling us to carry out the plans and projects which we intend to pursue—i.e., willpower.

Assuming that I can convincingly argue that adopting Self-Interest Theory would undermine the agent’s ability to exercise willpower, there are further implications. Of particular interest to me is the use of such a model of rational agency in the “law and economics” movement. This school of jurisprudence has had a very significant impact on such diverse and far-reaching topics as the theory of property rights, deregulation and the evolution of the common law. Of course, all of this bears significantly on the marketization of various national economies and of various domains of human life and interaction within a given economy. Much of the work of this school of thought presupposes a model of rational agency, which I would argue is not only inaccurate and tending to produce the kind of alienation that I discussed in my master’s thesis, but also tends to undermine the springs of action as they are embodied in willpower—thus disabling the agent from achieving the very kinds of maximization that economic rationality overtly endorses.

---

As yet, this is not so much the summary of an argument as the summary of a reasoned hunch. Under theories of willpower that I developed in earlier papers, but later discarded, the threat of Self-Interest Theory to willpower, via the connection of each to character, was easier to trace. I have continued to tinker with several different ideas as to how this linkage can be established, but I can afford to work none of them out in full until after this dissertation is completed. As they say: first things first.

Even if these latter paragraphs amount to little more than high-flown speculation for the moment, they do, I hope, at least breathe a little life into the suggested connection between willpower and higher-order theories that concern agents collectively by suggesting a particular route to establishing this linkage that, to this writer at least, may turn out to be of the greatest interest. But, ultimately, these are concerns for another day. The task at hand is large enough. However, before proceeding with the account of willpower, it is worthwhile to give the project a little historical context.

C. The Changing Fortunes of Willpower in Explanations of Behavior

Willpower has not figured large either in twentieth-century scientifically oriented accounts of human behavior or in twentieth-century analytic philosophy. Discarded as essentially a superstitious notion from an unscientific past, those who investigated human behavior sought motivational explanations elsewhere. In order to put this dissertation in context, I will briefly review the historical and philosophical reasons for the move away from using willpower in explaining human behavior, and briefly review the re-emergence of efforts to incorporate the concept of willpower into explanations of behavior.

---

The concept of the will as a causally potent force is an idea with a long philosophical history. It has been described as everything from the motive force issuing from the conclusions of practical reason, as, very roughly speaking, was the view of Aristotle, to a dynamic creative force underlying all reality, as seems, in simplest terms, to have been the view of Schopenhauer. The idea with which I am concerned is the idea of the will as a distinct motivational force in human action that persons employ voluntarily in a way that is not strictly constrained by the conclusions of reason as to what course of action is either most prudent or ethically best.

MacIntyre credits St. Augustine with first articulating an account of will as “the ultimate determinant of human action” which operates independently of both the conclusions of practical reason and “the passions.” The full-blown conception of the will as an independent power in motivation receives perhaps its most famous articulation by Descartes, who conceived of it as a special force operating outside of the laws of nature. This view flows from Descartes dualistic view of persons, as being comprised of two totally distinct substances: mind and body.

We are led, according to Descartes, to an agent that circumvents the hegemony of the mechanical-deterministic nexus by virtue of its immateriality and freedom. We are led to an agent whose behavior is governed by no other law than that which the agent himself creates. Descartes is explicit on this point. He says,

But the will is so free in its nature, that it can never be constrained. ... And the whole action of the soul consists in this, that solely because it desires

---


12 Alasdair MacIntyre, Whose Justice? at 156-57. See generally ibid., Ch. IX.
something, it causes a little gland to which it is closely united to move in a way requisite to produce the effect which relates to this desire.

On Descartes’ view the mind performs some form of psychokinesis every time a voluntary action occurs. It deploys its immaterial powers to move the body. The mind itself, however, is self-moved.\textsuperscript{13}

O’Shaughnessy argues that in the “most natural primitive philosophical conception of will” it is “a force whose power [is] inexplicable” and that is “subject to no internal constraint.” On such a view, individuals are credited with “harbouring limitless resources of will that might in principle be directed outwards.”\textsuperscript{14}

The cited passages highlight two of the more suspicious qualities of the will and its power when understood along Cartesian lines: (i) while not a part of the natural or material world it is causally potent, and (ii) it is “self-moving,” i.e. it is a cause arising from no antecedent causes, \textit{ex nihilo}, so to speak. Both of these alleged characteristics of will as a causal power have provided grounds for attacks by philosophers and scientists.

The concept of the will as a self-moving cause is of the most immediate importance to concerns about the explanation of human behavior in that it most directly relates to the causes of action. Further, it may survive as a theoretical explanation for action even after one has abandoned a robust Cartesian dualism. It is this idea of a self-moving cause that has been challenged as both conceptually incoherent and incompatible with any scientific explanation of human action.\textsuperscript{15}

Free will, in the traditional sense ..., is the power of agents to be the ultimate creators (or originators) and sustainers of their own ends or


purposes. This notion should be distinguished from free action, and not simply because free will is a power. To act freely is to be unhindered in the pursuit of your purposes (which are usually expressed by intentions); to will freely, in this traditional sense, is to be the ultimate creator (prime mover, so to speak) of your own purposes. Such a notion of ultimate creation of purposes is obscure, to be sure—many would say it is unintelligible—but there is little doubt that it is fueled by intuitions about free will from the beginning. Its meaning can be captured initially by an image: when we trace the causal or explanatory chains of action back to their sources in the purposes of free agents, these causal chains must come to an end and terminate in the willings (choices, decisions, or efforts) of the agents, which cause or bring about their purposes. If these willings were in turn caused by something else, so that the explanatory chains could be traced back further to heredity or environment, to God, or fate, then the ultimacy would not lie with the agents but with something else.

... 

The conception of free will ... as ultimate creation of purposes has been under attack since the seventeenth century for being obscure and unintelligible, and has been dismissed by many twentieth-century philosophers and scientists because of its supposed lack of fit with the modern image of human beings in the natural and social sciences.16

Thus, the concept of the will as a motive force operating outside of the causal nexus of the natural world seems to defy scientific account of human action in terms of natural laws which frame causal accounts in terms of heredity, environments or any other such factors.

Perhaps the most influential philosophical attack on willpower is that made by Gilbert Ryle in *The Concept of Mind*. Ryle’s arguments go to the heart of, and make plain, much of the critical thought or scientific instinct that bids one to abandon the concept of willpower. Specifically, Ryle attacks the notion that willpower is a specific causal mechanism that explains human behavior. Ryle argues that the idea that willpower

---

16 Kane, *The Significance of Free Will,* at 4-5.
as a causal force is merely an artifact of a bad philosophical theory. Ryle attacks this theory as follows:

According to the theory, the workings of the body are motions of matter in space. The causes of these motions must be either other motions of other matter in space or, in the privileged cases of human beings, thrusts of another kind. In some way that must forever remain a mystery, mental thrusts, which are not movements of matter in space, can cause the muscles to contract. To describe a man as intentionally pulling the trigger is to state that such a mental thrust did cause the contraction of the muscle in his finger. So the language of ‘volitions’ is the language of the para-mechanical theory of the mind.17

Ryle sees this conception of will as “just an inevitable extension of the myth of the ghost in the machine.”18

What Ryle means by “the para-mechanical theory of the mind” which is “just an inevitable extensions of the myth of the ghost in the machine” is expressed in its broad outline as follows:

The representation of a person as a ghost mysteriously ensconced in a machine derives from this argument. Because, as is true, a person’s thinking, feeling and purposive doing cannot be described solely in the idiom of physics, chemistry and physiology, therefore they must be described in counterpart idioms. As the human body is a complex organized unit, so the human mind must be another complex organized unit, though one made of a different sort of stuff and with a different sort of structure. Or, again, as the human body, like a parcel of matter, is a field of causes and effects, so the mind must be another field of causes and effects, though not (heaven be praised) mechanical causes and effects. ... 

... [S]ince mechanical laws explain movements in space as the effects of other movements in space, other laws must explain some of the non-spatial workings of minds.19

---


18 Ibid. at 63.

19 Ibid. at 18-19 (Italics in original).
The force of this reasoning, Ryle explains, leads the theorist to posit a theory of mind that is, in effect, a sort of negative image of the mechanical theory of bodies.

The workings of minds had to be described by the mere negatives of the specific descriptions given to bodies; they are not in space, they are not motions, they are not modifications of matter, they are not accessible to public observation. Minds are not bits of clockwork, they are just bits of not-clockwork.

... [T]hey are themselves just spectral machines.²⁰

Thus, in order to fill in the explanatory gap as to what sort of mental causes result in physical actions that are voluntary, such a theorist, so Ryle explains, relies upon the mysterious, para-mechanical “mental thrusts” which are dubbed “volitions.”

Ryle views the attempt to discern a causal explanation of the phenomena of will as a “category mistake,” leading us to look for the wrong kind of explanation because we have mistaken what class of phenomenon the phenomena of will are. More specifically, in this case it is purported to be a mistake which leads us to ask a causal question where it makes no sense to ask one, because the phenomena of will are not causal phenomena, and the concept of strength of will, as it is ordinarily used, is not a concept employed in trying to provide a causal explanation.²¹

Ryle’s attack on the idea of willpower was highly influential in both philosophy and the social sciences; however, social scientists were also faced with the immediate concern that they were aiming to develop empirically testable theories incompatible with the Cartesian conception of the will. Metcalfe and Mischel offer a succinct summary of the fall of willpower in psychological explanation:

²⁰ Ibid. at 19.

²¹ See Ibid. at e.g. Chapter I.
When psychology separated from philosophy at the end of the last century, theorists tried to leave behind the concept of willpower, but it repeatedly returns, cloaked in various metaphors. Freud... attempted to abandon it completely, making the issue of human volition moot by putting all the psychic determinants into unconscious motives; what we do, think, and feel is determined by the vicissitudes of biological impulses struggling for release from the id in its confrontations with the barriers and counterforces from the rest of the psyche. The intrapsychic wars were waged mostly beyond the individual’s awareness, in the continuous battles among an essentially hot id and cool ego and frigid superego. In this understandably pessimistic view of human nature and of the possibility—or impossibility—of volition that followed, Freud saw the only hope in efforts to replace the dark impulsiveness of the id with the reason and insights of the ego. ...

Unmoved by this metaphor..., the alternative theory that emerged, radical behaviorism, argued that behavior is under the control of external stimulus conditions, governed by the cues, contingencies, and consequences provided by the environment and the organism’s reinforcement history. Although they held opposite views of the locus of behavioral control, Freud and Skinner shared the conviction that willpower (or “personal agency” in current terminology) was just another fiction through which to delude ourselves.\(^\text{22}\)

Thus, willpower was ushered from the scene by scientific psychology.

This tendency to eliminate or define out of existence a robust, causal conception of willpower is still alive and well today. Kuhl has commented that care must be taken if we are not to yield to the temptation to prematurely eliminate such concepts from respectable scientific discourse.

Self-Regulation research is indeed beset with theoretical ambiguities to such an extent that many scholars have yielded to the temptation of “solving” all problems by a simple method: proclaiming volitional concepts as unscientific and reducing them to basic, nonvolitional concepts or even banning them from scientific psychology altogether. Despite the many conceptual confusions and operational problems involved in this research, I am convinced that psychologists will never be able to explain complex phenomena of human behavior without concepts like volition, willpower, self-regulation and self-control. I also believe

that such concepts will be again be neglected or banned by the scientific community unless we find better solutions to the theoretical and methodological problems raised by these concepts.\textsuperscript{23}

Given this general intellectual climate, it is not surprising that only seven years prior to the publication of Metcalfe and Mischel's paper reviewing the decline of volitional concepts in psychology, Ainslie published a book which purported to explain the phenomenon of willpower in which he states that "[t]he phenomenon of clear-eyed, wholly mental impulse avoidance described casually by such terms as 'resolutions' and 'willpower,' and often intended by the generic term 'self-control,' has not been explained by any other mechanism [aside from the one offered by Ainslie]."\textsuperscript{24} In fact, this remark of Ainslie's, with a few additional criteria added to the characterization of willpower (and a challenge to the account offered by Ainslie), sets the task for this dissertation.

Although, as I shall argue in more detail in what follows, Ainslie's may be the pioneering work in seeking to explain certain kinds of acts involving willpower, many others have joined in the effort to explain volitional concepts and concepts of self-control in recent decades. One article notes that "[r]esearch on self-regulatory issues has exploded since the 1960's."\textsuperscript{25} For example, recent attempts within psychology include the general multi-factor approach taken by Metcalfe and Mischel,\textsuperscript{26} and the account of "self-regulation" offered by Baumeister and Heatherton.\textsuperscript{27} Metcalfe and Mischel base their

\textsuperscript{23} Julius Kuhl, "Who Controls Whom When 'I Control Myself'?" \textit{Psychological Inquiry} vol. 7 no. 1 (1996): 61-68.


\textsuperscript{26} Metcalfe & Mischel "A Hot/Cool-System Analysis of Delay Gratification," at 3-19.

\textsuperscript{27} Baumeister and Heatherton, "Self-Regulation Failure: An Overview," at 1-15.
hopes for a theory of willpower on new advances largely based in cognitive psychology.\textsuperscript{28} By contrast, Ainslie looks to experimental studies of motivation in the behaviorist tradition and the application of microeconomic concepts to competing motivations within the agent.\textsuperscript{29} Even within the economic tradition, a theory of human behavior has been offered which defines willpower as an “effort variable” to be incorporated into its predictive theories, though the underlying mechanism is not itself explained.\textsuperscript{30} In philosophy, Bratman has offered an account of willpower in terms of his highly developed theory of intentions.\textsuperscript{31} Generally, the trend seems to have shifted toward more expansive understandings of the constituents of an explanation of human behavior and away from simplistic behaviorist accounts or scientifically unsophisticated Freudian accounts. These trends have fueled optimism about making willpower a respectable, or at least explicable, aspect of human behavior.\textsuperscript{32} However, before setting out the positive accounts of willpower, a task I reserve for succeeding chapters, there remain several tasks to perform in this chapter. As the first among these I will say more about what kind of explanation of willpower I plan to offer.

\begin{footnotesize}
\begin{itemize}
\item\textsuperscript{28} Metcalfe and Mischel, “A Hot/Cool-System Analysis of Delay Gratification,” at 3-19.
\item\textsuperscript{29} Ainslie, \textit{Picoeconomics} at 154.
\item\textsuperscript{32} In considering the broader historical development of ideas, O'Shaughnessy has offered a lengthy analysis that purports to explain why the phenomena of will should be of particular importance to philosophers and theorists of all kinds at the present time and why it is, in his phrase, “close to the nerve of the age.” As fascinating as this intricate explanation is, no short summary of two or three sentences can do it justice, and so the reader is left to explore and assess O'Shaughnessy's analysis for himself. Brian O'Shaughnessy, “Trying (As the Mental “Pineal Gland”),” \textit{Journal of Philosophy} vol. 70 no. 13 (July 1973): 365-86, Vol. I, Introduction.
\end{itemize}
\end{footnotesize}
III. The Explanatory Project

A. The Basic Strategy

In setting out the explanatory strategy of the project, the first order of business is to say, in a preliminary way, what I am seeking to explain. I have called it "willpower." What I intend to indicate by that term is, I hope, something familiar from everyday life. It is generally regarded as a resource we call upon when we seek to avoid succumbing to some temptation, or require some extra inner strength to carry out some difficult task. "[O]ur spontaneous way of conceiving the effort, under these circumstances, is as an active force adding its strength to that of the motives which ultimately prevail,"\(^{33}\) wrote William James at the end of the last century. James was not merely stating some view peculiar to the Victorian era.\(^{34}\) A century later, in our own times, another philosopher engaged in a detailed study of the will remarks that "... it is natural to think of 'the will' ... as the phenomenon of an action force in the mind: a psychic force that is exerted on (as impulsive act urge) and by (as striving) its owner."\(^{35}\) One theologian investigating the subject has remarked that the question "Does the will have 'power'? ...[ I]s for layman and expert alike the question to ask about willing."\(^{36}\)

In light of the history of the idea of will as a special motive force described in the previous section, this seems a large and daunting topic. However, something can be done

---


\(^{34}\) However, the concept may have had greater respectability among philosophers and psychologists of that time than it does today. *See Ainslie, Picoeconomics* at 154., § 5.3.4.


to prune it down to a more manageable size and make it a more plausible endeavor.

Bernard Williams has remarked:

… there is a deep difficulty about what an effort of will really is. There is indeed some kind of psychological process in connection with which the term is used—but might it not be just a psychological process which accompanied some actions (perhaps peculiarly difficult ones) and not others? Contrary to what some philosophers have supposed, efforts of will do not wear their metaphysical significance on their face. It may even be that they have none. 37

What Williams tentatively offers as a mere possibility is in fact my game plan. First, I am interested in uncovering the “psychological process” to which Williams refers. Second, I think it is a reasonable working hypothesis that whatever the process or processes involve, they accompany only certain difficult actions. Finally, it may be that this process has no metaphysical significance.

Taking these points in reverse order, I hope to offer an account of willpower that stands independent of any particular position in the great and longstanding metaphysical disputes concerning such issues as volitionism or the free will/determinism debate. I do not say that these metaphysical issues are unimportant, quite the contrary. It is simply that I am up to something else. However, if it turns out that a psychological explanation of willpower displaces a metaphysical treatment of some aspects of willing then so much the better at least to the extent that it makes the issues more tractable. This does not mean that there is no work left for the philosopher within this, but only that it is a different kind of work, namely, the philosophy of psychology and empirically-minded

theory of practical rationality. Moreover, the positive account of willpower that I will ultimately offer has its foundations, at least in part, in the work of philosophers: Elizabeth Anderson, J.L. Austin, Christine Korsgaard, Thomas Nagel, John Searle, and Bernard Williams among others.

Of course, all this depends crucially upon the next point—that is, on establishing that willpower is a phenomenon distinct from willing or volition generally and is not an accompaniment of all voluntary action. My means of doing this, following Williams’ suggestion, is first to describe actions undertaken by means of willpower as actions distinct from voluntary action generally, and second to account for the psychological processes or other factors giving rise to the features so described. There is some historical precedent for such a move in James’ treatment of the subject, as well as in that of another important theorist in this area, George Ainslie, and philosophers who have attempted to critique and build upon the work of Ainslie in this regard.

As I shall attempt to show in detail in the next chapter, James’ work in particular offers guidance to the present project in two ways. First, James offers a descriptive

---

38 For an example of this general approach to theory of practical rationality see David Pears, Motivated Irrationality, (South Bend, Ind.: St. Augustine’s Press, 1998). For a very different kind of example of the empirically-minded approach in philosophy of psychology/philosophy of mind see Daniel C. Dennett, Consciousness Explained, (Boston: Little, Brown and Company, 1991).

39 In this regard, this project is to be contrasted with works like that of O’Shaughnessy’s The Will: A Dual Aspect Theory, which seeks to explain willing as an element or aspect of all bodily actions. I am targeting the more specific concept of “willpower.”

40 James, The Principles of Psychology, Vol. II, Ch. xXVI.

41 See generally Ainslie Picoeconomics, Chs. 3-6, but see §§ 6.3 and 9.5 where Ainslie appears to argue that the same processes he employs to explain willpower underlie all choices experienced as voluntary. See also George Ainslie, and Nick Haslam, “Hyperbolic Discounting,” in Choice Over Time, eds. George Loewenstein and Jon Elster, (New York: Russell Sage Foundation. 1992a), 57-92.

characterization of willpower as it is ordinarily understood—the "folk psychological" theory of willpower, so to speak—that provides a solid target at which to aim our explanation, and which bears no necessary connection to the broader phenomena of willing. Second, James offers a psychological explanation of willpower that serves as an example of an explanation that answers to the observations grounding the folk theory of willpower by filling in an empirical psychological theory where metaphysical speculation used to stand. Interesting as James' positive theory and its modern descendants may be, they do not complete the work. For as I will argue, these theories leave unexplained the more difficult and interesting class of cases. Explaining them will occupy the balance of this dissertation.

In an article predating the current revival of interest in willpower, Henry Alker argued that inquiries into the nature or existence of willpower could take one of three forms, and that the three ways of asserting or denying the existence of willpower which these methods yielded mirrored the three ways of asserting or denying the existence of free will. These three methods are: (1) examining how the concept is used in ordinary language, (2) carrying out a scientific investigation to determine whether anything does or could correspond to the concept as it is ordinarily applied, and (3) demonstrating that the concept still has normative force and application, whatever the outcome of the scientific inquiry. From what I have already said, it is clear that this dissertation concerns the first and second methods of inquiry and proof. In Chapter Two I will first try to get clear on the target of the explanation, i.e. the phenomenon of willpower. Then I will try to explain it in the succeeding chapters. However, the inquiry

---

will not fall neatly within the more or less familiar lines of demarcation sketched by Alker, so I will need to say a bit more about the method of the project.

In trying to get a fix on willpower, I will not conduct any detailed linguistic analysis as Alker suggests. Rather, I will set out and argue for several features of what I take to be paradigm cases for a certain kind of exercise of willpower as supported by ordinary usage, introspection, phenomenological accounts and psychological experiments. After setting out the explanandum of the project in detail, I will proceed to examine numerous candidates, consisting of various motivational techniques and mechanisms which could be proposed as an explanation of willpower. Here, too, the account will diverge somewhat from what Alker appears to have in mind in that I am limited to philosophical, as opposed to scientific, methods. That is, I will only be able to offer an explanation of what could serve as a scientific explanation of willpower based upon the information currently available, rather than setting out directly to test a hypothesis as to what willpower is.

More generally, my explanation of willpower will be akin to explanations offered in anthropology. On at least some views, part of the anthropologists’ job consists of two parts. The first part is to study what is sometimes called the “emic” conceptions or belief systems of the subject population, i.e. those belief systems that they hold as a part of their ordinary inculturated understanding of the world. The second part is to test or explain those beliefs in terms of an “etic” theory based upon what are regarded as objective and scientific concepts and methodologies. 44 Again, this is only an analogy since I rely neither on my own fieldwork nor on my own scientific experiments, but rather upon the

observations of others and my own efforts to account for them within a consistent theory. At the same time, I intend to offer a theory which can be supported or falsified, at least in part, by empirical testing and observation broadly construed.

Nonetheless, pursuing the anthropological analogy a little further, the emic analysis concerns the folk psychology of willpower. What I mean by folk psychology is the conceptions of ordinary persons about how and why beliefs are held, actions are taken, and so on. This is not to say that all agents universally have such conceptions, for it may be that persons differ in fundamental ways in their conceptions of self, motivation and action across cultures.\footnote{See e.g. Clifford Geertz, "‘From the Native’s Point of View’: On the Nature of Anthropological Understanding," in \textit{Local Knowledge: Further Essays in Interpretive Anthropology}, (New York: Basic Books, 1983a), 55-72; Clifford Geertz. “The Impact of the Concept of Culture on the Concept of Man,” in \textit{The Interpretation of Cultures}, (New York: Basic Books, 1973a): 33-54.} I am principally concerned with agents operating in the cultures of the West since the late 19\textsuperscript{th} century – though my analysis may be applicable to agents living in the same era as Descartes and perhaps even that of St. Augustine.

Further, it must be borne in mind that even where such conceptions are generally shared within a culture during a given period, individuals may vary in their beliefs and understandings.

In focusing on willpower, I am examining a phenomenon more complex than those with which philosophers in the longstanding debates about folk psychology are generally concerned, namely beliefs and desires. Roughly speaking, the philosophical debate about folk psychological concepts primarily concerns whether folk-psychological posits such as "beliefs" and "desires" correspond to anything real on the best or most scientific understanding of human thought and behavior. The view that they do not, and
are not a part of any proper scientific theory has been called "eliminitivism." This is not a battle I choose to join in this dissertation. However, this dissertation does presuppose that some form of elementary folk-psychological explanation is useful in accounting for human behavior, since the analyses and explanations that I offer are couched in terms of the elementary terms of folk psychology. My account is built up in terms of components such as beliefs and desires—at least as these terms have been employed in their most general sense in the behavioral sciences.

Since the folk theory of willpower will turn out to look much like the kind of philosophical theory attacked by Ryle, it is worth pausing a moment to make a brief preliminary response to his criticisms. As I see it, to a certain extent, Ryle simply misses the mark. For what I am out to explain is the folk concept of willpower, not an "artificial" philosophical theory about volition. I will try to show that the power of the will as it is conceived in folk psychology is a causal force. It is counterproductive to waylay the investigation by erecting a straw man called "volition" and then argue that it has no application in the pure "ordinary language" of the ordinary folk, and, hence, is a mere philosophical fantasy. Ordinary people do talk of exercising willpower, and that is not the fault of philosophers. Even if it were, the concept has already gone so far in melding into and influencing ordinary language concepts that it is no longer separable from ordinary language. The deficiency in the characterization of the "ordinary" use of

\[\text{Footnote: For a good discussion of this debate see Stephen Stich, From Folk Psychology to Cognitive Science: The Case Against Belief, (Cambridge: the MIT Press, 1983):}.\]
the concept of "willpower" which Ryle offers in his efforts to explain away the reputed causal attributes of willpower will be touched upon in the next chapter.

The real challenge in light of Ryle's criticism is to see whether the commonly understood causal view of willpower can be made respectable without resort to "ghostly thrusts" dressed up in respectable philosophical jargon. That, indeed, is one of the larger purposes of this dissertation: to offer an account of willpower without resort to "ghostly thrusts" that arise ex nihilo and tip the balance in the agent's motivational system and allow the agent to avoid succumbing to temptation. Even if there is no correlate in reality to the folk concept of willpower in all its dimensions, it will be the work of this dissertation to show that there is a psychological mechanism that exhibits the core functional features associated with the folk-concept.

The etic theory that I propose is reductionist in that it breaks down willpower into component parts—but it stops well short of eliminativism. In this regard it is probably not too dissimilar from most contemporary accounts of human action either in terms of their overt ontology or their tacit ontology, which may be obscured by technical jargon. However, at some points, my explanation will end in mere comparison with other familiar psychological phenomena. That is, as some points the explanation will come to an end merely by pointing out that there is a commonly accepted psychological phenomenon which shares a particular functional feature with willpower as I have explained it. The implication is that if we credit the existence of a functional feature as quite ordinary in one context, we ought not to balk at it as being mysterious and unscientific in another. It may be that a further explanation is possible for both of the compared phenomena, but at present such explanation, if there is one, is outside the scope
of this dissertation. At other points, I will be able to take the analysis further, using philosophical concepts and analysis that provide a grounding for some of our social scientific observations. In sum, the account will have both vertical and horizontal aspects, but will leave most elementary folk psychology intact—although willpower itself will be shown not to live up to some of the magical qualities sometimes ascribed to it by folk psychology, at least not in a literal sense.

Due care must be taken in offering such an explanation. For as Alker reminds us, it is open to the proponent of willpower, as the term is ordinarily used, to argue that the concept does not contain the kind of empirical theory or assumptions that one may try to ascribe to it, and so isn’t open to being discredited at the hands of science. The view that I will ultimately argue for lies between two opposite poles. I will argue that certain uses of the word “willpower” do imply certain commitments about entities and causal relations in the world, but I will argue that these commitments need not imply belief in the more magical properties of willpower, though in many cases such beliefs might be present.

Further, in seeking an explanation of willpower, I will be looking for something that serves the functional role ascribed to willpower in its ordinary use. That is, when a person is said to be using willpower *something* is happening. I will argue for a mechanism or technique that functions much as willpower is reputed to on the folk conception. Hence, I argue that an agent’s reliance upon willpower is not like wishing upon a star. There is something real behind the concept; the challenge is to say what it is. In sum, I hope to do as much as possible to account for the folk concept without simply
dismissing it or explaining it away while at the same time remaining true to the 
explanatory constraints of science.

B. Basic Motivational Assumptions

If willpower is to be understood as a causally potent force without resort to the 
"ghostly thrusts" ridiculed by Ryle, it must be explained within a larger theory that 
presupposes causal regularity in motivation. Further, there are a number of other basic 
motivational assumptions that will be common to most paradigms for describing and 
explaining human action causally. I will rely openly on some of these assumptions in this 
dissertation. I will now briefly review these assumptions and define the motivational 
terminology to be used in this dissertation.

First, action is goal oriented. That is, action is undertaken to achieve some end. I 
will take no stance as to what sorts of entities serve as ends—e.g. whether they are 
events, states of affairs or some other ontological category or combination of categories. 
By agent's "end" I mean an actual or potential object of action without specifying 
whether that object has what I will call "reward value" for the agent. Of course, in 
specific instances, the context will make clear that in a given example under discussion a 
given end does have reward value for an agent. Where an end does have reward value 
for an agent, I will sometimes make use of another behaviorist term and refer to it as a 
"reward," intending that the term be understood in the broadest possible sense as 
encompassing whatever the ends of human action may be.

Second, motivation is presumed to operate in a causally regular manner. The 
"reward value" of some event or state of affairs is the degree to which it disposes that
subject to emit behavior that produces that event. Thus, by definition, the greater a reward, the more the subject will be disposed to seek it. Thus, for purposes of this dissertation, the “reward value” of a given reward and the “strength of the agent’s motivation” to obtain that reward are one and the same. Thus, relative to some set of sufficiently specified background conditions for a specified agent a given reward should have one and only one reward value. That is to say motivation is presumed to operate in a causally regular manner. That is not to say that conditions can always be reproduced by researchers to allow for meaningful comparisons between competing rewards for one agent over time, nor for comparisons of relative reward value for a given end between agents. However, much of the fruitful work done in the human sciences suggests that such comparisons are sometimes possible.

I will put forth no specific theory of the nature of motivation. To speak of the agent’s “motivation” to do x implies that doing x or obtaining x has reward value for him. In other words, all things being equal he will try to obtain x. Ordinary language or various psychological, philosophical or economic theories of action or behavior may have more specific commitments as to the nature of motivations, but no such commitment is necessary here. I have no need to postulate, for instance, that the agent is motivated to obtain his ends because they produce in him a pleasurable mental state, or that a motivation is explicable in terms of some specific physiological phenomenon. My interests in motivation only concern some of its formal features. I will make clear what these formal features are in the remainder of this section. In this way, the arguments

---

47 See Ainslie, *Picoeconomics*.
presented in this dissertation can engage the broadest possible spectrum of theories of human action.

This is a good point at which to stop and get clear on some terminology. From time to time I will write of an agent’s “desire.” In such instances, “desire” should be understood as a synonym for “motivation” unless the contrary is clearly indicated by the context. Finally, by saying that an agent has a “preference” for some end over another, I mean that the former has a greater associated reward value than the latter. Some caution must be taken here not to confuse the notion of “preference” that expresses a relation between reward types and one that merely expresses a relation between two particular reward tokens. I will try to make clear in all succeeding uses of this word which way I am using it.

Third, typically the agent will act in pursuit of the reward with the greatest reward value amongst the competing options. This is not to say that an agent will always do so. Psychological, physiological, physical or other constraints may mean that the agent will not always act in pursuit of that end which under ideal circumstances he would pursue. However, it follows from the manner in which “reward” and “reward value” have been defined that the greater the reward value, the more the agent is inclined to pursue it. Although I will not attempt to present or justify some underlying theory of utility, or general theory of human decision, I think it is safe to say that most empirically oriented theories of behavior presuppose that generally the agent is able to make some kind of comparison between competing rewards and that the resulting decision and behavior will typically involve seeking that end which has the greatest reward value relative to other ends. In other words, at least ideally, a rough and ready preference
structure for an agent could be worked out, and it would generally be predictive of behavior—or at least have sufficient predictive value to be of interest to behavioral scientists.

An important qualifier to this manner of discussing motivation is that tomorrow’s (or yesterday’s) motivations have no effect upon today’s actions. That is, I have explained a reward in terms of the behavioral response it will produce in the agent. The fact that tomorrow the agent will be in some motivational state wherein, if presented with the opportunity to obtain some reward, say, the chance to have a steak dinner, he will gladly expend effort or resources to get it, does not imply that he will be motivated to pursue that end today. He may be so motivated or he may not be, that question cannot be answered except by gaining an understanding of his motivations at the present point in time. In other words, “reward value,” as I have defined it, is specified relative to background conditions that change over time, making “reward value,” in practical terms, a time-relative concept.

This raises several issues about how the agent is, or should be, motivated in the present, given the fact that at some later point in time he will come to have a new desire for some reward, or some existing desire will become greatly strengthened relative to existing rewards. More specifically, an agent may know that he will have such desires in the future, and also know that he will not be able to satisfy them without taking some action today. Ordinarily, we want to say that the agent has some reason to take the necessary preparatory actions, and we expect him to be motivated to do so. If, for instance, I know that tomorrow I will want to have pasta for dinner, but will not have time to go to the grocery store, we would say that I have a reason to go to the grocery
store to buy pasta today. We would also expect that I would have the motivation to go to
the grocery store today.

How it happens that an agent comes to be motivated in the present to satisfy his
future (and not yet present) motivations is an interesting and important question,\(^{48}\) but
one beyond the scope of this dissertation. That agents often do is a commonplace of
human experience. But however this kind of motivational phenomenon occurs for
agents, I have defined the concept of “reward value” in a such a broad and general way
that work can proceed while remaining neutral on this point. For however the objects of
future motivations come to result in present actions, by virtue of the definition of “reward
value” and “motivation,” we can say that where they do there must be some current
motivation that generates the action without explaining either the nature of that
motivation or how it comes to be that the agent has it.

Further, what does or does not constitute a reason for carrying out actions in the
present in order to obtain the objects of future (though not yet present) desires, or, for that
matter, what constitutes a reason for carrying out any action, are further issues that I can
gladly leave to one side. An account of reasons will not be necessary either to
demonstrate the flaws in existing accounts of behavioral phenomena that serve as
candidates for the causal/scientific explanation of willpower, nor to put forward a
positive account of this phenomenon within that framework. Explaining the nature of
reasons for action is a large and important task for theorists of human action, but
concerns fundamental aspects of decision-making and action not at issue here. My aim is
to explain a specific phenomenon of behavior within a broadly accepted framework,

rather than to challenge, support or explain the fundamental aspects of that framework or of decision-making or action in general.

Fourth, I will assume that a preference remains sufficiently constant over time to be a useful concept in the explanation and prediction of behavior. That is, in order for “reward value” to be a useful concept in the explanation of behavior, the previously referenced background conditions relative to which the reward value is fixed must be sufficiently broad or enduring or recur with sufficient frequency that meaningful predictions can be made. If, for instance, reward value were fixed only in relation to background conditions which occurred only momentarily and were never again repeated, then a measure or estimate of the reward value of some end \( x \) would provide no information as to whether the agent would be likely to pursue \( x \) to the exclusion of some other end, \( y \), in the future.

Ordinarily, we expect some motivations to endure through changing circumstances—such as the desire to practice a certain profession. Similarly, we expect some desires to recur over time, such as the desire for chocolate ice cream. We also expect these desires to maintain their reward value over time, at least relative to a sufficiently broad class of competing rewards, absent some special intervening circumstances requiring explanation. For instance, we would expect that where a person prefers practicing medicine over practicing law that this preference will persist absent some intervening circumstances. If the preference changes, we would tend to say that this fact requires some explanation. Similarly, if someone typically prefers chocolate to vanilla ice cream, the next time a desire for ice cream arises within him, we would expect the desire for chocolate ice cream to be greater than the desire for vanilla. If it were not,
we would say that this change requires some explanation. The question as to why many preferences remain constant in this way is beyond the scope of this dissertation. However, the fact that they do is a presumption common to causal explanations of action, and a presumption relied upon in this dissertation.

Fifth, in a sufficient number of cases to be behaviorally significant, reward value may be aggregated. By aggregation, I do not mean that reward value must be purely additive in all the relevant cases. Rather, a much less contentious assumption may be made. All that is required is the assumption that where \( x \) and \( y \) each have significant reward value for an agent, an agent would be significantly more motivated to take some action that will result in his obtaining \( x \) and \( y \), than he would be where that action would only result in his obtaining \( x \).

It is not necessary to the argument of this dissertation that the reward value of any two rewards may be aggregated in this way, but merely that in a sufficient number of cases to be significant in the explanation of behavior, the reward value of ends is will be subject to aggregation. In many cases aggregation will not hold. For instance, I may be motivated in some degree to go to the refrigerator to get a glass of soda. I may also be motivated to go to the refrigerator to get a glass of milk. But my overall motivational strength will not be the sum of these two motivations for the simple reason that obtaining one of the rewards eliminates the appeal of the other. In other words, I don’t want to drink both milk and soda. However, in many instances, reward value can be aggregated. If, for instance, I can obtain both a newspaper and a soda at the store, and I want both of these items, I will be more motivated to go to the store than if I could only obtain one of
these items there. Again, this is in line with common experience, as well as most or all causal/scientific explanations of behavior, and is not expected to be a point of contention.

In the discussions which follow, points one through five above will frame the explanations offered, examined and criticized. I will refer to points one through five collectively as the "Basic Motivational Assumptions," or "BMA" for short.

IV. Conclusion

I have tried to indicate, in a preliminary way, what the target of my explanation is and to show why this concept of willpower is interesting to both philosophers and scientists interested in explaining human behavior or addressing practical concerns arising from human behavior or motivation. I have also tried to show the historical decline of this concept in scientifically respectable explanations of behavior as well as the resurgence of interest in the phenomena of will as ineliminable and necessary elements of the explanation of action. I also described in general terms how I hope to explain willpower as a scientifically acceptable mechanism that causally influences human behavior and that answers to the folk description of willpower. Finally, I set out the basic motivational assumptions which serve as the background for such a theory of willpower. Having done this preliminary work, I can move on in the next chapter to give a more detailed account of what it is that I am trying to explain, and what an explanation of it would look like.
Chapter Two
Folk and Scientific Theories of Willpower

I. Chapter Introduction

In this chapter, I will use William James' *The Principles of Psychology* both to make a beginning on the exegesis of the folk concept of willpower and to use James as a sort of philosophical foil in order to clarify the target concept of willpower. Along the way, I will argue that James and his theoretical heirs have done significant work in explaining one of at least two existing folk concepts of willpower. In so doing, they offer examples of how empirical theories may be used to capture specific psychological processes that have been described as the action of willpower. However, both James and later theorists have largely neglected to explain another distinct concept of willpower, even though James himself discusses it at some length. This second concept of willpower has largely been swept under the theoretical carpet because it seems to require the kind of magical volitions criticized in Chapter One. By the end of this chapter, I will have given a fairly clear picture of this more problematic concept of willpower and its neglect in contemporary psychological theory. I will conclude by drawing together and summing up the features of that kind of willpower since an explanation of this concept is the aim of my theory of willpower.

II. James on the Folk Theory of Willpower

A. The Phenomenology of Willpower

James argues that there is a class of decisions to take action that involve a special kind of exertion that we commonly associate with willpower. Most decisions are made without such a sense of effort, emerging sooner or later on the basis of a rational
assessment of the evidence, a particularly apt conceptualization of the situation, circumstance, accident, natural inclination, impulse, change of mood, or some other such event, circumstance or activity. Such is not the case with decisions involving willpower.

James describes this special case as follows:

In the ... final type of decision, the feeling that the evidence is all in, and that reason has balanced the books, may be either present or absent. But in either case we feel, in deciding, as if we ourselves by our own willful act inclined the beam; in the former case by adding living effort to the weight of the logical reason which, taken alone, seems powerless to make the act discharge; in the latter by a kind of creative contribution of something instead of reason which does reasons work. The slow dead heave of the will that is felt in these instances makes them a class altogether different from the... preceding classes [of decision]. What the heave of the will betokens metaphysically, what the effort might lead us to infer about a will-power distinct from motives, are not matters which concern us yet. Subjectively and phenomenally, the feeling of effort, absent from the former decisions, accompanies these. Whether it be the dreary resignation for the sake of austere and naked duty all sorts of rich mundane delights, or whether it be the heavy resolve that of two mutually exclusive trains of future fact, both sweet and good, and with no strictly objective or imperative principle of choice between them, one shall forevermore become impossible, while the other shall become reality, it is a desolate and acrid sort of act, an excursion into a lonesome moral wilderness. If examined closely, its chief difference from the three former cases appears to be that in those cases the mind at the moment of deciding on the triumphant alternative dropped the other wholly or nearly out of sight, whereas here both alternatives are steadily held in view, and in the very act of murdering the vanquished possibility the chooser realizes how much in that instant he is making himself loose. It is deliberately driving a thorn into one's flesh; and the sense of inward effort with which the act is accompanied is an element which sets... [this] type of decision in strong contrast with the previous[ly] discussed ... varieties, and makes of it an altogether peculiar sort of phenomenon.¹

Such is James' characterization of the phenomenology of the exercise of willpower.

B. The Folk Theory of Willpower

Beyond the mere description of the experiences falling under the folk-psychological heading “exercising willpower,” James also provides what we might think of as a rudimentary folk-psychological theory of willpower. James writes:

Now our spontaneous way of conceiving the effort, under all ... [the various circumstances James describes], is as an active force adding its strength to that of the motives which prevail. When outer forces impinge upon a body, we say the resultant motion is in the line of least resistance, or of greatest traction. But it is a curious fact that our spontaneous language never speaks of volition with effort in that way. Of course if we proceed a priori and define the line of lease resistance as the line that is followed, the physical law must also hold good in the mental sphere. But we feel, in all hard cases of volition, as if the line of coarse motivation were the more pervious and easy one, even at the moment when we refuse to follow it. He who under the surgeon’s knife represses cries of pain, or he who exposes himself to social obloquy for duty’s sake, feels as if he were following the line of greatest temporary resistance. He speaks of conquering and overcoming his impulses and temptations. ...

The ideal impulse appears, in comparison with ... ['lower’ propensities], a still small voice which must be artificially reinforced to prevail. Effort is what reinforces it, making things seems as if, while the force of propensity were essentially a fixed quantity, the ideal force might be of various amount.... And if a brief definition of ideal moral action were required, none could be given which would better fit the appearance than this: It is action in the line of greatest resistance.

The facts may be most briefly symbolized thus, P standing for the propensity, I for the ideal impulse, and E for the effort:

I per se < P.

I + E > P.

In other words, if E adds itself to I, P immediately offers the least resistance, and motion occurs in spite of it.

But the E does not seem to form an integral part of the I. It appears adventitious and indeterminate in advance. We can make more or less as we please, and if we make enough we can convert the greatest mental
resistance into the least. Such, at least, is the impression which the facts spontaneously produce on us ... 2

Such, then, is the folk theory of willpower as James understands it.

C. Initial Criticisms of James’ Account of the Phenomenology and Folk Theory of Willpower

Rich as James’ descriptions of the phenomenology and folk theory of willpower are, I propose two caveats. First, though somewhat ambiguous, James appears at some points to speak of willpower in terms of its role in decisions to act as opposed to its role in motivating the actions themselves. I think this is a mistake. There is a real sense in which decision alone is not enough, at least in some cases. Consider the decision to dive into a cold swimming pool and “get it over with,” as opposed to slowly inching one’s way in and prolonging the discomfort. Assume here that one dreads the cold water and there is a great temptation to try, however vainly, to put off the full intensity of the discomfort that immersing oneself in it will involve by inching one’s way in slowly. One may have this temptation even with the full knowledge that, when all is said and done, the total of discomfort experienced by inching one’s way in is greater than the total of discomfort experienced in simply diving in. In such a case one could say: “The decision about what to do is easy, it’s diving in that is hard.” From this, we might draw some evidence that the added motivation at issue in willpower propels the action rather than the decision.

A second, though connected, concern about James’ characterization of the phenomenon of willpower is his characterization of the “effort” involved. Again, James is ambiguous. However, James generally describes the feeling of effort as connected with the decision to act, rather than with the action undertaken. Caution should be

2 Ibid. at 548-49 (emphasis in original).
exercised here as well, for James will ultimately argue that the only effort involved is the effort of focusing on the idea of the action to be undertaken. This, too, is a mistake. Consider again our would-be swimmer. It seems reasonable to say that at least part of the “effort” relevant to his exercise of willpower in diving into the pool is that involved in the physical action of running toward the pool and jumping in. I will not speculate on just what exactly produces this sense of effort. I simply want to show that there is some plausibility to the claim that the sense of effort attaches to the action undertaken in the exercise of willpower, in addition to, or, perhaps, instead of, the mere decision to act.

One might object that my example misses the mark because a decision to act without the concomitant action is not a “real” decision, or is an empty decision, or is deficient in some other way however one may choose to characterize it. Regardless of how one frames this response to my objection, I must reply that I do not intend to stake out a position in the analysis of the linkage between decisions and actions. I wish only to make the point that on a very natural understanding of willpower, it involves not merely decision, but also action. In what follows, I will attempt to explain the biases in James’ characterization of willpower by making the case that James’ description of the folk concept of willpower is prejudiced by the psychological theory he is advancing.

Regardless of the outcome of the subsequent arguments, for the moment I have achieved my purpose. That is, I have set out James’ phenomenology of willpower together with his rendition of the folk theory of willpower. Of course, theory may invade any attempted description of the experience of exercising willpower to a greater or lesser degree; and in this case, the folk theory may be little more than some generalizations about the phenomenology of willpower. Nonetheless, I have tried to separate out in a

40
rough and ready way what goes under the heading of theory, and what under the heading of observation, for one operating within the folk theory itself. I then proposed two amendments to James’ view by suggesting that (i) willpower attaches to actions rather than (only) decisions, and (ii) the feeling of effort associated with willpower is connected, at least in part, with actions taken and not solely with decisions to take action. Subject to these amendments, James’ account of willpower on the folk conception provides a good starting point in identifying the target of the explanation of willpower presented in this dissertation.

III. James’ Scientific Theories of Willpower

A. Introduction

The next order of business is to set out James’ scientific theories of willpower. (As I will set out below, James has two distinct theories of willpower.) This will allow me to achieve several aims. First, an assessment of James’ theories of willpower will help to sharpen the focus on the specific features of those behavioral performances that are the exercise of willpower. Second, it will reveal how James’ biases—created by James’ commitments to certain psychological theories—may have prejudiced his phenomenology of willpower causing him to misdescribe the operation of willpower in the two ways discussed in the preceding section. Third, the assessment of James’ theories will show them to be deficient in crucial respects, which both clears the field for further advances in the pursuit of an explanation and hints at direction that such explanation will have to go. Fourth, rightly understood, James may be taken as providing an example of how one can make a successful non-metaphysical explanation of willpower without crossing over into the territory claimed by those who would offer explanations of the will generally or take a stand in the free will/determinism debate.
This fourth point involves establishing that James offered an account of what one may choose to call a kind or an aspect of willpower—even though he fails to account for the kind of phenomenon that he actually describes so lucidly before beginning his scientific explanations. In short, it turns out there are at least two kinds of willpower. With these aims in mind I proceed to consider James’ scientific account of willpower.

R. James’ Primary Theory of Willpower
To begin, I must set out two fundamental principles of James’ scientific psychology. According to the first, a person is naturally motivated to take an action if the idea of that action is present in his mind. James writes:

We may lay it down as certain that every representation of a movement awakens in some degree the actual movement which is its object; and awakens it in a maximum degree whenever it is not kept from so doing by an antagonistic representation present simultaneously in the mind.\(^3\)

Further, these ideas have motivational power according to the following principle:

The states of mind which normally possess the most impulsive quality are either those which represent objects of passion, appetite, or emotion—objects of instinctive reaction, in short; or they are feelings of ideas of pleasure or of pain; or ideas which for any reason we have grown accustomed to obey so that the habit of reacting on them is ingrained; or finally, in comparison with ideas of remoter objects, they are ideas of objects present or near in space and time. Compared with these various objects, all far-off considerations, all highly abstract conceptions, unaccustomed reasons, and motives foreign to the instinctive history of the race, have little or no power.\(^4\)

According to these two principles, of the various courses of action entertained by the agent, that which is represented by the idea with the “most impulsive quality,” wins out. Thus, for ordinary voluntary action, there is no special sense of effort, because the action follows naturally and automatically from the idea of its object, which serves to

---

\(^3\) Ibid. at 526.

\(^4\) Ibid. at 536.
activate it. Of course, such ideas may be framed through a process of deliberation or affected by causal factors not associated with the will so that these influences bear upon how the ideas figure in the process of decision.\(^6\)

With respect to the special operation of willpower, James’ basic view is that it amounts to focusing the attention on the better or more prudent end of action. James writes:

\[\ldots \text{[W]e see that attention with effort is all that any case of volition implies. The essential achievement of the will, in short, when it is most ‘voluntary,’ is to attend to a difficult object and hold it fast before the mind. The so doing is the fiat; and it is a mere physiological incident that when the object is thus attended to, immediate motor consequences should ensue.}\(^7\]\n
Or again:

This strain of the attention is the fundamental act of will. And the will’s work is in most cases practically ended when the bare presence to our thought of the naturally unwelcome object has been secured. For the mysterious tie between the thought and the motor centers next comes into play, and, in a way which we cannot even guess at, the obedience of the bodily organs follows as a matter of course.\(^8\)

Thus, according to James, does willpower operate.

However, getting the idea of the prudent course of action into one’s mind is not enough; it must displace the competing ideas. In the first instance, under James’ general principles of connection between ideas and action, the tempting idea or thought of action that one would seek to overcome would almost always tend to have the “most impulsive

\(^5\) Ibid. at 522-28.

\(^6\) Ibid. at 528-535.

\(^7\) Ibid. at 561.

\(^8\) Ibid. at 564.
quality.” Hence, if both alternatives are held steadily in view, the impulsive alternative, i.e. the temptation, would tend to win out. However, James argues that the idea associated with the more prudent or reasonable course of action has the power to dispel the more passionate or immediate idea:

... by a sort of self-preserving instinct which our passion has, it feels that these chill objects, if they but once gain a lodgement, will work and work until they have frozen the very vital spark from out of all our mood and brought our airy castles in ruin to the ground. Such is the inevitable effect of reasonable ideas over others—if they can once get a quiet hearing ....

Sustained in this way by a resolute effort of attention, the difficult object ere long begins to call up its own conjurers and associates and ends by changing the disposition of the man’s consciousness altogether. And with his consciousness, his action changes, for the new object, once stably in possession of the field of his thoughts, infallibly produces motor effects. The difficulty lies in sustaining the field. Though the spontaneous drift of thought is all the other way, the attention must be kept strained on that one object until at last it grows, so as to maintain itself before the mind with ease. ...

Consent to the idea’s undivided presence, this is effort’s sole achievement. Its only function is to get this feeling of consent into the mind. And for this there is but one way. The idea to be consented to must be kept from flickering and going out. It must be held steadily before the mind until it fills the mind. Such filling of the mind by an idea, with its congruous associates, is consent to the idea and to the fact which the idea represents. 

In sum, James’ hypothesis is that, once attended to and purposely held in attention, the “rational” idea will drive its more impulsive rival from the field, thus precipitating “rational” action. I will call this hypothesis as a proposed explanation of willpower: “James’ Primary Theory of Willpower,” or “JPTW.”

---

9 Ibid. at 536.

10 Ibid. at 563-64 (emphasis in original).
C. Problems with James' Primary Theory of Willpower

Regardless of our assessment of the truth or accuracy of JPTW with respect to thought and action more generally, it faces several glaring difficulties as an explanation of the specific phenomenon of willpower. The most obvious problem is that it contradicts James' folk or introspective description of willpower, which depicted a class of decisions where:

... its chief difference from the ... [other] cases appears to be that in those cases the mind at the moment of deciding on the triumphant alternative dropped the other one wholly or nearly out of sight, whereas here both alternatives are steadily held in view. ¹¹

Yet on JPTW, when the more prudent idea is "stably in possession of the field" and "fills" the mind, it must, in whole or in part, displace the more impulsive idea. This is what allows "rational" or more abstract ideas to win out over ideas that naturally have a more "impulsive quality" in the competition to govern behavior. Hence, JPTW necessarily hinges upon violating one of the descriptive criteria of willpower. The objection that JPTW, or any other such scientific theory of willpower, conflicts with the descriptive criterion that "both alternatives are steadily held in view," that one remain, so to speak, "clear-eyed" ¹² in recognizing exactly what is at stake as the decision then poses itself, I will call (with apologies for the awkward expression) the "Clear Eyedness Objection." James recognizes this problem, and so advances a second hypothesis. But, before getting to this, it is worthwhile to stop and consider other shortcomings of JPTW.

¹¹ Ibid. at 534.

Another apparent shortcoming of JPTW is that it fails to meet another of James' descriptive features of willpower: that of a new motivation entering into the factors bearing upon action. As already noted, James states that it is "our spontaneous way of conceiving the effort [of will]" in thinking of it "as an active force adding its strength to the motives which prevail."\textsuperscript{13} Spelled out more fully, the "[effort of will] does not seem to form an integral part of the ... [ideal impulse]. It appears adventitious ..."\textsuperscript{14} Support is lent to this idea by the fact that the effort of will seems "indeterminate in advance," such that "[w]e can make more or less as we please, and if we make enough we can convert the greatest mental resistance into the least."\textsuperscript{15} JPTW leaves no room for this aspect of willpower, since focusing on the rational or prudent idea draws forth not some new motivation, but merely enhances the motivations associated with that idea in the first place. Thus, for these reasons, JPTW fails to account for a phenomenon that answers to the description that James himself gives of the phenomenology and folk theory of willpower. I will call the objection that a theory such as JPTW fails to provide an account of the special motive involved in willpower, but rather relies upon some apparently ordinary motive for the action, the "\textit{Ordinary Motivation Objection}.

JPTW is also subject to the objection that it fails to reflect the proper connection between decision and action. Here the discussion of this point is more important for my purposes than arriving at a definite conclusion with respect to how JPTW, as it stands or in modified form, ultimately fails in the face of this objection. In the first instance, it

\begin{itemize}
\item \textsuperscript{13} James, \textit{The Principles of Psychology}, Vol. II, 548.
\item \textsuperscript{14} Ibid. at 549.
\item \textsuperscript{15} Ibid. at 548-49 (emphasis in original).
\end{itemize}
looks as if James’ Primary Theory makes willpower, in effect, an effort to set loose an automatism. There is a two-step process:

Step 1: attention fixing on an idea;

which is followed by:

Step 2: the idea produces a corresponding action through an automatic process.

Since Step 2 is an automatism, the objection goes, the process just amounts to deliberately setting loose an automatism, and this can’t be willpower. Rather, it is the instrumental employment of a mechanism or device.

It is as if one were unable to raise one’s leg voluntarily, so one strikes oneself below the knee thereby triggering the patellar reflex and causing the leg to rise. Such a strategy may give one the added “power” to raise one’s leg, and one may have voluntarily caused the leg to rise, but one has not thereby voluntarily raised his leg. Efforts motivated by willpower should be efforts undertaken voluntarily. In fact, willpower should be part of the voluntary motivation operative in producing the action. Since the motivation operative in producing the action as described in James’ Primary Theory is merely an automatic response, the activity described cannot result from the operation of willpower.

James’ first line of defense is likely to be that Step 2 is an element of all voluntary actions. Hence, any attack on James’ Primary Theory on these grounds implies an attack on his general theory of voluntary action. Such an undertaking is beyond the scope of this dissertation. More importantly, any sustained defense of James’ theory of voluntary action will be dependent upon a more general premise about motivation. That

---

16 Ibid. at 522-569.
is: all action is goal directed, and there is a fixed causal relationship between the object of
the action and the presence and strength of the corresponding motivation to obtain it in a
given agent at a given time. This principle may be cashed out in a number of ways
depending upon the psychological, economic or other theory espoused. Regardless of
how one chooses to articulate it in a specific theory, it is a basic premise of the scientific
explanation of behavior. This is a point that I highlighted in Chapter One as a part of
BMA. For now, the important thing is that this is a premise which I cannot challenge
since it is implicit in any scientific account of willpower, and producing such an account
is, after all, the aim of this dissertation.

The objection, however, can be reformulated. The real issue is not so much
whether the action at issue is produced by means of an automatism, but whether,
generally speaking, agents operating on the basis of the folk theory of willpower would
perceive it as such so that it undermines the viability of the folk theory. For such a
perception would strongly militate against such an agent calling the behavioral
phenomenon at issue an exercise of willpower. This is an interesting issue, but one, I am
afraid, that takes me too far afield into the psychology of voluntary action. Instead, I will
focus on a related, though more narrow and easily grasped, consideration.

Related to the topic of voluntariness, though separable from it, is the question of
the connection between the agent’s choice and the operative motivation. Specifically, the
issue is whether the motivation is intrinsically connected with the decision to act, or
whether it is produced through some separate process. An example illustrates the point.
Consider again the swimmer. Suppose he simply says to himself, “Oh how I hate cold
water. But I am just going to do it!” and then jumps in. So far, the swimmer’s behavior
sounds like it could be an instance of exercising willpower. Imagine a second case. The swimmer steps out from under the sun umbrella and stands in the sun. He does this with the intent of making himself uncomfortably hot, knowing that this will increase his desire to simply jump into the pool and get some immediate relief. With this added motivation, he will have sufficient “nerve” to simply jump right in.

In both cases, the decision to jump in was obviously voluntary, and the motivation to do so in both cases apparently arose from a voluntary decision—perhaps most clearly so in the second case. Yet there is something suspect in the second decision. Set aside for the moment that the motivation seems to be of the wrong kind for an instance of willpower—i.e. it is not a special motivation distinct from the primary desires being weighed by the agent. The important point is that in the second case the agent’s motivation to act was clearly created not by a decision to take action, but by a decision to stimulate the motivation to take action. In other words, the agent engaged in a bit of self-manipulation in order to get himself to act. Although the swimmer’s action did not follow automatically—in the manner of a full-blown reflex—from the intentionally stimulated motivation, this added motivation appears to have been necessary to tip the balance in favor of diving in. The fact that the added motivation was achieved through a secondary process, by means of a “device,” so to speak, rather than as a corollary to, or a part of, the primary decision to act, makes it suspect. The precise connection between decision and willpower is difficult to specify, and I will not attempt it here. I will flesh out this connection further in the course of this dissertation. What is clear for now is that

\[17\] The use of this particular word to get at this idea was suggested by a remark made by Mary Warnock. D.F. Pears, J.F. Thomson and Mary Warnock, “What is the Will?” in Freedom and the Will, ed. D.F. Pears, (London: Macmillian, 1965): 14-37 at 35.
in the cases of willpower with which we are concerned the motivation must arise directly
in connection with the decision to act, rather than through a separate process, event or
decision such that once set loose it then impinges upon the decision to act. I will refer to
this as the "Device Objection." I will call the general problem of specifying the proper
connection between decision and willpower the "Linkage Problem."

It is not altogether clear whether this objection is genuinely telling against James'
Primary Theory. The kind of focusing on our aims that he argues stimulates motivation
may be a natural part of many performances classed amongst efforts of will. Considering
the case of one trying to get out of bed, Warnock states that:

... it does not seem to be just a contingent fact that this act of will involves
concentrating on the idea of getting up [out of bed]. In other words, the
point that ... an effort of will consists in concentrating one's attention on
the idea of the action is a necessary truth."

Even if we doubt the correctness of this claim, it is at least some evidence that it may
often be a natural part of performances regarded as efforts of will that the agent focuses
intently on the aim of the action he hopes to undertake. If this conclusion is correct,
then in at least some instances agents may not see such focusing as a device employed
separately from an effort of will. Such performances could then be classed as the
exercise of willpower by agents operating within the folk concept of willpower.

Thus far I have raised several objections to James' Primary Theory of Willpower.
Specifically, I have shown JPTW to be subject to the Clear-Eyedness Objection and the
Ordinary Motive Objection, and to not be obviously out of the woods with respect to the

---

18 Ibid. at 34-35.

19 It may also be that Warnock, and the other participants in the discussion in which she makes this
comment, were influenced by James' theories. James is referenced several times in the text, and some of
the views expressed bear a remarkable resemblance to those of James. Ibid. at 14-37.
Device Objection. Regarding the Clear-Eyedness Objection in particular, James realized that JPTW was in some trouble. However, the game is not entirely over for James; he has a second theory of willpower.

D. James’ Supplementary Theory of Willpower

Recognizing that JPTW is subject to the Clear-Eyedness Objection, James advances a second, supplementary theory of willpower. Softening the unavoidable objection to JPTW, James frames it in this way:

It is unqualifiedly true that if any thought do fill the mind exclusively, such filling is consent. The thought, for that time at any rate, carries the man and his will with it. But it is not true that the thought need fill the mind exclusively for consent to be there; for we often consent to things whilst thinking of other things, even of hostile things; and we saw in fact that precisely what distinguishes our ‘fifth type’ of decision [i.e. the form of decision associated with willpower] from the other types is just this coexistence with the triumphant thought of other thoughts which would inhibit it but for the effort which makes it prevail. The effort to attend is therefore only a part of what the word “will” covers. ... ²⁰

This amounts to an admission that James has offered no account of willpower as he himself has described it phenomenologically. For James’ “fifth type” of decision is the very type he associates with willpower, as I explained earlier.

What, then, does James have offer to cover this class of cases? Very little: the word ‘will’ ... covers also the effort to consent to something to which our attention is not quite complete. Often, when an object has gained our attention exclusively, and its motor results are just on the point of setting in, it seems as if the sense of their imminent irrevocability were enough of itself to start up the inhibitory ideas and make us pause. Then we need a new stroke of effort to break down the sudden hesitation which seizes upon us, and to persevere. So that although attention is the first and fundamental thing in volition, express consent to the reality of what is attended to is often an additional and quite distinct phenomenon involved.

The reader’s own consciousness tells him of course just what these words of mine denote. And I freely confess that I am impotent to carry the analysis of the matter any farther, or to explain in other terms of what this consent consists. It seems a subjective experience *sui generis*, which we can designate but not define. We stand here exactly where we did in the case of belief. When an idea stings us in a certain way, makes as it were a certain electric connection with our self, we believe that it *is* a reality. When it stings us another way, makes another connection with our Self, we say, *let it be* a reality. To the word ‘is’ and the words ‘let it be’ there correspond attitudes of consciousness which it is vain to seek to explain. The indicative and the imperative moods are a much ultimate categories of thinking as they are of grammar. The ‘quality of reality’ which these moods attach to things is not like other qualities. It is a relation to our life. It means our adoption of things, our caring for them, our standing by them. This at least is what it practically means for us; what it may mean beyond that we do not know. And the transition from merely considering an object as possible, to deciding or willing it to be real; the change from the fluctuating to the stable personal attitude concerning it; from the ‘don’t care’ state of mind to that in which ‘we mean business,’ is one of the most familiar things in life. We can partly enumerate its conditions; and we can partly trace its consequences, especially the momentous one that when the mental object is a movement of our own body, it realizes itself outwardly when the mental change in question has occurred. But the change itself as a subjective phenomenon is something which we can translate into no simpler terms.21

That is the whole of what I will call “James’ Supplementary Theory of Willpower,” or “JSTW,” as James presents it in The Principles of Psychology.

E. Problems with James’ Supplementary Theory of Willpower

The explanation that James offers in support of this supplementary theory seems an amalgam of vestigial elements of his primary theory and hand-waving toward something more. Instances of the former include the introduction of the scenario wherein the agent first causes the pertinent idea to exclusively dominate his consciousness thereby bringing himself to the brink of appropriate action when “the sense of [their] imminent irrevocability” of such action causes “inhibitory ideas” to rise up and interrupt the move

21 Ibid. at 568-69 (emphasis in original).
from the endorsed idea of rational action to the act itself.\textsuperscript{22} This seems an effort to draw limits around, and even to shrink, the territory where the more difficult-to-explain features of willpower are operative. However, it seems that willpower operates in many psychological contexts other than the one so described by James, and in any event James’ hypothesis does nothing to take the mystery out of just how the more intractable features of willpower do operate when they do come into play.

Similarly, talk of an idea which “stings us in a certain way” or makes “a certain a certain electric connection with our self” seems an effort to harken back to the theory that the “impulsive quality” of the idea impels action.\textsuperscript{23} But, \textit{ex hypothesi}, it is the idea with the more impulsive quality that is to be overcome in the case of willpower. Further, whatever this supposed quality of the idea is that leads us to say “let it be a reality,” it cannot arise from the endorsement of reason—i.e., the conclusion that \textit{this} is the course that it best or most rational. For decisions arising from such a conclusion, on James’ view, involve an entirely separate class of cases.\textsuperscript{24} In his original phenomenological descriptions of willpower quoted above, James described willpower as coming into play where reason does not finally determine which action is to be taken.\textsuperscript{25} James’ talk of an idea that makes a certain “electric connection” seems to amount to nothing more than the hypothesis that there must be some quality “\textit{x}”, which some ideas have, that (i) lead us to consent to those ideas that have it and (ii) cannot otherwise be identified, and (iii) is somehow outside the normal scale of “impulsive quality.” It seems that (ii) sounds the

\textsuperscript{22} Ibid. at 568.

\textsuperscript{23} Ibid. at 568-69.

\textsuperscript{24} Ibid. at 531-35.

\textsuperscript{25} Ibid. at 534. \textit{See generally} Ibid. at 531-35.
death knell for any hope of a testable theory along these lines. Still, (iii) may be most troubling in that there is no explanation as to why, if such an idea is capable of this “electric connection” that impels action, it does not simply dominate the field as against the common run of impulsive ideas in the first instance.\textsuperscript{26} If, in the end, action springs from this “electric” quality of the idea representing the action, why would one need willpower at all? It seems one is back to the normal and natural dominance of some ideas over others by virtue of their “impulsive quality.” No special effort of willpower is needed. In short, this aspect of James’ Supplementary Theory of Willpower, to the extent that the explanation moves from the quality of the idea to the action the idea propels, is at best completely \textit{ad hoc} and utterly unfalsifiable, and at worst, in outright contradiction with the background theory that it presupposes.

As I already stated, there is also the hand-waving aspect of JSTW. This is where, for instance, James references a change in “attitude” toward the ends of action or our ideas of them. James gestures at the operative attitude as “caring for” or “standing by” the thing, or “saying” “let it be a reality” or having a “we mean business” attitude.\textsuperscript{27} James states that “this phenomenon is something which we can translate into no simpler terms,” but calls upon the reader to attend to his own consciousness as it will “tell him of course just what these words of … [James] denote.”\textsuperscript{28}

James’ characterization is helpful in that it is suggestive of certain characteristics of willpower, but I think he is mistaken in two key respects. First, the more typical

\textsuperscript{26} Ibid. at 568-69.

\textsuperscript{27} Ibid. at 569.

\textsuperscript{28} Ibid. at 568-69.
“attitude” in cases of willpower is not one of “consent,” not a saying to oneself: “let it be,” but rather the affirmative: “I will do it.” By this latter expression I do not intend a prediction about the future movements of one’s body, but rather an assertion that one intends to take action—it involves a kind of commitment. This, I would suggest, is the kind of resolution we can most easily imagine our hypothetical swimmer making before he takes the plunge.

Confirmation of these views is found in the experimental studies of Ach:

Experimentation on willing has proved to be difficult. Yet it has been attempted with some enlightenment by N. Ach and others of the so-called Wuerzburg School. ... In most of his studies, Ach asked subjects first to learn combinations of nonsense syllables and then ordered them, in the experimental phase, to respond with something different from the response learned. In other words, he established willing as an act going against formed habits (analogous to Kant’s ethical stipulation of willing as going against spontaneous pleasure). From his observations and introspective and retrospective accounts of his subjects he developed the following purely descriptive facts:

In such acts of willing there is a subjective feeling of strong tension, generally accompanied by body sensations especially of skeletal musculature: there is (literally) a “jerk” of the will.

There is an awareness of intentionality, usually accompanied by an image of the goal and mostly and act of inner speech: “I must ...” “now I have to ...”, “I am going to ...” This is usually a formulation of a future action in which the person is not yet fully engaged. It is often accompanied by a feeling of power or capacity: “I can. ...”

There is a high awareness of actuality, an experience of “I will” with emphasis on the I that does the willing. “When I seriously will, I also can” or “When I don’t take this seriously, something else will happen.” It implies a definite not-willing of other things, and this is again an experience in which the ego feels itself taking the lead and assuming responsibility.

There is a consciousness of goal and instruments. This contains the object of the act of willing set for oneself and the means necessary for its attainment. In Ach’s studies these goals and instruments may be merely schematic, not visually concrete.
There is a consciousness of effort and difficulty.  

Ach’s observations nicely confirm James’ introspective account of willpower. However, Ach’s characterization of the experienced inner resolve in terms of a first-person resolution to take future action is, I think, a more accurate phenomenological description than James’ “let it be” as described in JSTW. The “inner speech” or, to foreshadow what is to follow, “speech act” is one of commitment rather than assent. Of course, all this requires more explanation, and I intend to give such an explanation in Chapter Four. But for now I can say that this points out another respect in which I take issue with JSTW: I think there is more, much more, that can be said about the operative “attitude” involved in willpower and its connection with action than James indicates.  

F. The Feeling of Effort

Ach’s observations also highlight the importance of the “feeling of effort” associated with willpower. This is certainly an important phenomenological aspect of the cases of use of willpower with which I am concerned. However, I have little to contribute by way of an explanation of this feature of willpower. I think that the test of whether my account of willpower properly meets this criterion is not one that can be answered in this dissertation since it would ultimately require verification through experimental studies and observation, subject interviews and the like.

---


30 As will become clear in Chapter Four, I think James points in the right direction for the start of that explanation in referencing the “categories” of grammar that parallel the “categories of thinking.” Ibid. at 569. However, James never follows up on these ideas in The Principles of Psychology.
In fact, James provides what amounts to a long and fascinating cautionary tale in the history of attempts to explain such feelings of effort in relation to phenomena of will.\textsuperscript{31} The principle object of James’ attack is the view that, in the case of a voluntary act, “[t]here must … be a special current of energy going out from the brain into the appropriate muscles during the act ….”\textsuperscript{32} In the view of the theorists supporting this position, this “special current of energy” would produce a “feeling of innervation” which they associated with volitional effort.

In the course of attacking this view, James offers numerous clinical examples through which the supposed feeling of innervation/effort is explained (or, more accurately, explained away) in terms of ordinary muscular activity. In one particularly instructive passage, James quotes Wundt, an “innervation” theorist, and then goes on to offer a critique of Wundt’s views based on the work of Ferrier:

Wundt says that were our motor feelings of an afferent nature [i.e. carrying nerve impulses from the muscles back to the nerve center],

“it ought to be expected that they would increase and diminish with the amount of outer or inner work actually effected in the contraction. This, however, is not the case, but the strength of the motor sensation is purely proportional to the strength of the impulse to movement, which starts from the central organ innervating motor nerves. This may be proved by observations made by physicians in cases of morbid alteration of muscular effect. A patient whose arm or leg is half paralyzed, so that he can only move the limb with great effort, has a distinct feeling of this effort: the limb seems to him heavier than before, appearing as if weighted with lead; he has, therefore, a sense of more work effected than formerly, and yet the effected work is either the same or even less. Only he

\textsuperscript{31} James, \textit{The Principles of Psychology}, Vol. II, 486-522.

\textsuperscript{32} Ibid. at 493.
must, to get even this effect, exert a stronger innervation, a
stronger motor impulse, than formerly.”

In complete paralysis, also, patients will be conscious of putting forth the
greatest exertion to move a limb which remains absolutely still upon the
bed, and from which no afferent muscular or other feelings can come.

But Dr. Ferrier in his Functions of the Brain . . . disposes very easily of
this line of argument. He says:

“It is necessary, however, to exclude movements altogether
before such an explanation [as Wundt’s] can be adopted.
Now, though the hemiplegic patient cannot move his
paralyzed limb, though he is conscious of trying hard, yet
he will be found to be making powerful muscular exertion
of some kind. Vulpian has called attention to the fact, and I
have repeatedly verified it, that when a hemiplegic patient
is desired to close his paralyzed fist, in his endeavors to do
so he unconsciously performs this action with the sound
one. It is, in fact, almost impossible to exclude such a
source of complication, and unless this is taken into
account very erroneous conclusions as to the cause of the
sense of effort may be drawn. In the fact of muscular
contraction and the concomitant centripetal impressions,
even though the action is not such as is desired, the
conditions of the consciousness of effort exist without our
being obliged to regard it as depending on central
innervation or outgoing currents.

“It is, however, easy to make an experiment of a simple
nature which will satisfactorily account for the sense of
effort, even when these unconscious contractions of the
other side, such as hemiplegics make, are entirely excluded.

“If the reader will extend his right arm and hold his
forefinger in the position required for pulling the trigger of
a pistol, he may without actually moving his finger, but
simply making believe, experience a consciousness of
energy put forth. Here, then, is a clear case of
consciousness of energy without actual contraction either
of the one hand or the other, and without perceptible bodily
strain. If the reader will again perform the experiment, and
pay careful attention to the condition of respiration, he will
observe that his consciousness of his efforts coincides with
a fixation of the muscles of his chest, and that in proportion
to the amount of energy he feels he is putting forth, he is
keeping his glottis closed and actively contracting his
respiratory muscles. Let him place his finger as before, and continue breathing all the time, and he will find that however much he may direct his attention to his finger, he will experience not the slightest trace of consciousness of effort until he has actually moved the finger itself, and then it is referred locally to the muscles in action. It is only when essential and ever-present respiratory factor is, as it has been, overlooked, that the consciousness of effort can with any degree of plausibility be described to the outgoing current. In the contraction of respiratory muscles there are the necessary conditions of centripetal impressions, and these are capable of originating the general sense of effort. When these active efforts are withheld, no consciousness of effort ever arises, except in so far as it is conditioned by the local contraction of the group of muscles towards which the attention is directed, or by other muscular contractions called unconsciously into play in the attempt.

"I am unable to find a single case of consciousness of effort which is not explicable in one or the other of the ways specified. In all instances the consciousness of effort is conditioned by the actual fact of muscular contraction. That it is dependent on centripetal impressions generated by the act of contraction, I have already endeavored to show. When the paths of the centripetal impressions or cerebral centers of the same are destroyed, there is no vestige of a muscular sense. That the central organs for the apprehension of the impression originating from muscular contraction are different from those which send out the motor impulses, has already been established. But when Wundt argues that this cannot be so, because the sensation would always keep pace with the energy of muscular contraction, he overlooks the important factor of the fixation of the respiratory muscles, which is the basis of the general sense of effort in all its varying degrees."33

Thus, James takes himself to have refuted Wundt’s innervation theory.

Whatever the ultimate merit of the James-Ferrier view of the feeling of effort in voluntary bodily movement, the quoted passage does establish several important points.

First, the source of the phenomenological feeling of effort may not always be apparent

33 Ibid. at 503-05 (quoting (1) Wundt, Vorlesungen über Menschen und Thierseele, I. 222. and (2) Ferrier, Functions of the Brain (Am. Ed. 222-24)).
even to the agent carrying out the action. Particularly if we are under the sway of some theory, whether scientific or philosophical, we may be inclined to think that such feeling must arise from some special exertion or energy or mysterious volition, rather than from some collection of several quite ordinary causes.\textsuperscript{34} Hence, the second point: James’ explanation of the source of the sense of effort suggests that extreme caution should be used in trying to determine the origin of the feeling of effort through theoretical or conceptual argument rather than careful empirical study.

Despite these lessons to be learned from James’ own work, James’ theoretical commitment to JPTW and the motivational theory on which it rests seems to push him to conclude that the “strain of the attention is the fundamental act of will” and “the immediate point of application of the volitional effort lies exclusively in the mental world.”\textsuperscript{35} Interesting as this idea may be, given what has already been said about the failings of JPTW, effort of attention cannot be the full explanation of the sense of effort in all cases where willpower is applied. Further, it seems a commonplace that the kinds of effort and associated fatigue involved in efforts of will may involve various kinds of mental, emotional and physical strain, typically present in combination. More than this I am reluctant to say, for as I have already noted, the best way to gain an understanding this particular issue in the explanation of willpower seems to reside in careful

\textsuperscript{34} Interestingly, O’Shaughnessy, in his monograph on will, which is devoted primarily to will as it is involved in bodily action, concludes that the related notion of “trying” “proves to be nothing but the immediate active event effect of a desire to act.” Brian O’Shaughnessy, The Will: A Dual Aspect Theory, 2 vols. (Cambridge: Cambridge University Press, 1980), Vol. II, 115. Without going into the details of O’Shaughnessy’s complex treatment of the subject, I merely intend to note that his characterization of “trying” with respect to bodily action appears to dovetail nicely with the idea that the sense of effort involved in trying, in the case of bodily action, is to be found in the muscular action itself—i.e. the typical “immediate active event effect of a desire to act,” rather than some special sensation of the effort of volition.

\textsuperscript{35} James, The Principles of Psychology, Vol. II, 564.
experimental observation and subject interviews, rather than armchair theorizing or
introspection.

G. The Effect of James' Scientific Theories on His Description of the Folk
Concept of Willpower
At this point I return to my initial criticisms of James' depiction of the
phenomenology and folk theory of willpower, i.e. that it exhibits a bias toward describing
the effort involved as an effort of decision rather than and effort of action, and similarly,
that it tends to see the special motivation involved as impinging upon decision rather than
upon action. We can see why, despite the fact that there is good reason to see the
motivation involved in the class of cases of willpower with which I am concerned as
bearing upon primarily upon actions, rather than decisions, such a description must have
seemed to James to have been mistaken or simply out of court. Given James' hypothesis
of the fixed causal law operative between ideas and action, if an additional motivation
were to be operative in behavior it would have to factor into behavior prior to setting this
causal process in motion. That is, such an added motivation would have to bear upon the
forces allowing one to hold the action-triggering idea in mind. James characterizes this
process of fixing the idea of the action one will take in one's mind as a kind of decision
to take action.\textsuperscript{36} Thus, James' scientific theories of motivation and of willpower pushed
him to describe willpower as a phenomenon of decision rather than of action even under
the folk account.

By the same token, the forgoing observations on the feeling of effort, together
with the previous sections setting out the relevant principles of James' scientific
psychology, make clear how the basic principles of James' psychological theories tend to

\textsuperscript{36} Ibid. at 531-35.
bias his description of willpower in describing the direction of the effort associated with willpower as involving the decision to act as opposed to the action itself. For on James’ view, an action follows automatically from the idea of the action held in consciousness. Therefore, any unique kind of effort involved in cases of willpower must be outside this relation. Since this is James’ universal principle of motivation for actions, there can be no secondary motivation of willpower operating in tandem with it. Accordingly, with no parallel source of motivation operative, there is nothing for any special sense of “effort” to be associated with as concerns the action itself. Hence, on JPTW, willpower can only be the mental operation of fixing in the mind that idea which generates the motivation to action. The only place where a special sense of effort fits into such a theory is in connection with the effort to fix the operative idea in the mind which James takes to be a part of the decision to take action. Thus James’ scientific theory of willpower pushes him to describe willpower as bearing upon decision rather than action, and to describe the corresponding sense of effort as being a result of the focusing of attention that is a factor in the decision as opposed to a motivation propelling the action itself. As I have already argued, this seems to misdescribe the situation from the standpoint of the phenomenology of willpower and the basic tenets of the folk-psychological theory.

My treatment of James has not yet concluded, it is not yet time to summarize my conclusions about what his theories of willpower have to contribute to this project. Nonetheless, for the sake of clarity I will draw a line under what I have done to this point and sum things up. Thus far, I have reviewed James’ rendition of the folk theory of willpower, his two scientific theories of willpower, and problems with both his scientific and folk accounts of willpower. Before finally concluding my review of James’ work, it
will be helpful to see how James’ work can be better understood as an effort to explain only one of at least two concepts of willpower. Additionally, I will provide some examples of how more recent theorists have followed James’ lead in attempting theories of the less controversial class of cases involving willpower, leaving the field wide open for the challenge of coming to grips with the more difficult cases. It is the latter challenge that I will take up in the succeeding chapters.

IV. Two Concepts of Willpower

Does James’ Primary Theory of Willpower do nothing to explain willpower? Actually, I think it does a great deal—or at least it points in the direction of theories that could do a great deal. The problem is that James’ explanation is not really aimed at the phenomenon that he ostensibly sets out to explain. However, what he in fact provides an explanation of could also be described as a class of cases involving willpower. That is, there are arguably at least two classes of cases of willpower, even under the folk conception. James has just set out to explain one after first stating that he intended to explain the other.

JPTW describes cases where, by focusing on one’s endorsed aim, one causes an increase in the motivation to obtain that aim. Consider again the swimmer. We could easily imagine a case where the swimmer does focus on how refreshing it will feel paddling around in the cool waters once he has gotten used to it. By so focusing his attention, he may well be able to generate sufficient motivation to dive in. I have no objection to calling this a case of willpower. As I have already remarked, the act of focusing attention on the intended action or its consequences and the resultant increase in the motivation to carry out that action is a natural one within the folk concept of
willpower. The only point I am trying to make here is that this operation of willpower fits neither with James phenomenology of willpower nor with the description of it implicit in the folk theory that he articulates. I have already set forth my reasons for this claim in Section III.

What I hope to do is to offer a scientifically acceptable explanation of willpower that is consistent with James' description of the phenomenology and folk theory of willpower. I have highlighted some of the features of this class of cases in Sections II and III, focusing on features such as the sense that a special added motivation is operative, the observation that there is a certain direct connection between a resolution to act and the increase in motivation, the apparent lack of reliance upon the technique of focusing attention on the endorsed aim at the expense of attentional focus on the tempting aim, and a number of other features, some of which have only been briefly mentioned. Again, in such a case our hypothetical swimmer may simply say to himself, "I am going to do it!" and then take a running leap into the water without any apparent change in his relative attentional focus as to the anticipated unpleasantness of the jolt of coldness upon impacting the water or the relief he will experience after becoming accustomed to the water.

These latter features of willpower all hang together naturally in James' phenomenology and his rendition of the folk theory of willpower. I think James is right in his description. I also think that this is the more challenging class of cases to explain and one that is central to motivating the folk belief in the metaphysical will that operates as an uncaused cause of indeterminate strength. In order to get clearer on just what the target of the explanation is, in Section VII I will sum up and enumerate the features of

---

37 See supra Section III.C.
this class of cases of the exercise of willpower, and will also offer a more detailed and robust description of what I take to be its key features.

I will call the kinds of behavioral performance upon which I will focus my positive explanation “Output Willpower.” The reason for this is that these behavioral performances seem to involve at their core the sense of some new and special added motivation being introduced to the motivational system that is conceived of as the will. The agent has a sense that his self or soul or mind has generated the causal force of will—that it has generated some new force that enters into the motivational mix. This new force, the will, is conceived of as the operative output of the agent’s resolution to act.

By contrast, cases that fit with the attentional focus model seem to concern what is allowed into the motivational hopper by controlling what is allowed to fall within attentional focus. Both the motivations generated by the objects of attentional focus and the end result of the competition between them are regarded as more or less fixed and natural results of the relevant objects having entered into the agent’s field of consciousness to a greater or lesser degree. If the agent is to exercise any control over his motivations to take action it must occur indirectly and at an earlier stage. He must control which motivations are generated, or at least their relative strength, by dint of controlling what he focuses his attention on. All else follows automatically from that point. In sum, the point of control is on the input side—on what is allowed into consciousness so as to generate motivations. Hence, I call instances of this kind of behavior performance instances of “Input Willpower.”

---

Obviously, there will be some similarities between cases of Input Willpower and Output Willpower. For instance, both are means of overcoming temptation, both may probably be employed with greater or lesser force or power in a given instance, and so forth. My principle aim in this chapter is to sufficiently distinguish the two classes of cases so as to motivate my positive project. In this section, I have also tried to show how this distinction allows us to reaffirm the importance of the work James has done in the explanation of willpower while recognizing the large gap remaining in the explanation of certain aspects of this phenomenon.

V. Theories of Input Willpower

The successors to the basic strategy of James’ Primary Theory of Willpower in particular, and of the effort to explain Input Willpower in general, are many and varied, often relying upon a multifactor approach. For instance, Metcalfe and Mischel hypothesize a general distinction between a “cool system” of motivation and a “hot system” of motivation within the agent.\(^{39}\) In the most general terms, Metcalfe and Mischel hypothesize that the greater the extent to which stimuli are processed through the hot system and activate hot “nodes,” the greater the likelihood that they will lead to immediate, cognitively unmediated action, while the converse is true for stimuli processed through the cool system.\(^{40}\) Thus control strategies include “obscuring” or “covering” the tempting object, avoiding giving attention to the tempting object, having

---

241-260 at 242-43. Hoffman makes a similar distinction with some interesting historical background. However, at the end of the day we are dividing up the array of cases in significantly different ways. Ultimately, nothing important hangs on the terminology used, it is the description and categorization of cases that matters.


\(^{40}\) Ibid. at 5-9.
other stimuli present which distract attention from the tempting object, directing one’s attention to thoughts of other unrelated goals or rewards, having the tempting object presented in a less immediate or tangible form, and thinking of the tempting object in a more “cool” or less immediate or tangible form.⁴¹ This theory seems to curiously echo James’ notion of motivating ideas that have various degrees of “impulsive quality.”

Variations on the theme of attention control are numerous. Some theorists have studied in detail how the ways in which temptations are attended to or imagined and how attentional efforts interact with various kinds of plans for action may be used by persons seeking to direct their efforts away from the tempting object.⁴² Others have focused their studies on the comparative effects of, on the one hand, imagining successful outcomes or, on the other, imagining taking the actions necessary to produce those successful outcomes.⁴³ Some theorists have focused upon the motivational effects of thinking of actions in terms of their higher-level meanings within the larger context of goals and plans, and alternatively, of focusing on the lower-level aspects of action that come closer to the mere physical elements of the action itself and, perhaps, its most immediate aim.⁴⁴

Baumeister and Heatherton focus upon a normally operative self-regulatory system that involves: standards, the monitoring of standards, and an “operate” phase

---

⁴¹ Ibid. at 9-13.


wherein steps are taken in light of the results yielded by monitoring. Baumeister and Hetherton make much of the role of attention control in the operation of such systems. While these authors make much of a “strength model” of self-regulation in that such abilities involve limited resources which may be taxed with use, and describe how the successful operation of the system often involves choice, nothing akin to a clearly defined conception of Output Willpower appears to be present in their system. Further, Baumeister and Heatherton present these ideas in a “target article” with commentary offered by numerous other contributors. Despite their varied theoretical perspectives, none of these commentators, with the exception of Ainslie, describes or suggests any


46 Ibid. at 4-6.

47 Ibid. at 2-4, 6-9.

explanation for a clearly defined concept of Output Willpower. (I will consider Ainslie’s theory in detail later.)

In what amounts to a kind of control-of-attention theory, Locke and Kristof create a curious amalgam by combining Ayn Rand’s theory of the freedom of the will with Lock and Latham’s theory of goal setting to reach the ultimate conclusion that:

Willpower is usually viewed as a separate faculty (the “will”) that directly overwhelsoms or outmuscles reason and/or conflicting desires in order to attain some end. In contrast, we view reason as the source of the will. We would view willpower as the choice to use one’s rational faculty to select goals, to identify the means to achieve them, and to keep one’s goals and the reasons for planning and pursuing them in focal awareness at the time of action in the face of contrary thoughts and emotions. Willpower is volition (focus) put to use—that is, put into action.49

Thus, Locke and Kristof explicitly abandon Output Willpower altogether in favor of a concept of volition which amounts to careful attention to practical reasoning and persistent attention to goals formulated in light of one’s practical reasoning.

The tendency to simply rule out of existence by definition any concept of Output Willpower as Locke and Kristof have done, and as Ryle did before them, is one I am trying to combat. In this vein, it is worth noting Shapiro’s criticism of Baumeister and Heatherton’s article. Shapiro notes a tension in Baumeister and Heatherton’s article that I have found to be common, in one form or another, in many, if not almost all, such theories and which is related to one of the criticisms I made of JPTW:

The authors rely heavily on the concept of self-regulation strength for their analysis of self-control: “If the impulses have strength, whatever stifles them must presumably consist of some greater strength.” This concept of self-regulatory strength seems in some ways to be a surrogate for a concept of volition, but nowhere is its nature actually made clear. ...

The authors believe that self-regulation or self-control involves an "overriding" of impulses; however, overriding requires strength. Strength involves the capacity to delay gratification. Finally, self-regulation sometimes requires transcendence (of immediate temptation) but transcendence is a form of "attention control." In all this . . . there is a circularity; we have come very little further than the old-fashioned, quasimoral faculty of willpower. Like willpower, these concepts are presented in an implicitly voluntaristic way: One must "delay gratification," "forego" temptation, and "mentally transform" boring tasks into interesting ones by focusing awareness beyond the immediate stimuli. But this will not do; volitional control or regulation cannot be explained by invoking voluntaristic constituents. Altogether, it seems to me that dealing with the problems of self-regulatory failure, as they are called here, requires a clearer picture of action or self-direction.\footnote{Shapiro, "The 'Self-Control' Muddle," at 77. See Baumeister and Heatherton, "Self-Regulation Failure: Past, Present and Future," at 93-96 for a response.}

I think Shapiro is right in his basic point that a lot of "scientific" theories purporting to explain willpower or related concepts first involve redescribing and relabelling the issue (i.e. "self-regulation" as opposed to "willpower" or "self-control") and then offering a more or less complex and scientific-sounding account which ultimately falls back on assumptions or metaphors reliant upon voluntaristic concepts that the subject theory has officially disclaimed (and often purports to explain away in reductionist terms).

As I have tried to explain, my general strategy is to avoid this problem by, first, having a very restricted target. That is, I seek only to explain the nature and functioning of the added motivation cases of Output Willpower while leaving untouched the logically and causally prior concepts involved in volitions and voluntary actions generally. Second, my explanation of Output Willpower will do away with voluntary concepts except that of voluntary choice—and as to the explanation of that concept I remain agnostic.
In general then, it seems fair to say that the theories surveyed in this section are typical of the scientific explanation of willpower. Like James’ Primary Theory of Willpower, they have something important to say about motivational processes, and may help to explain the mechanisms underlying the folk concept of Input Willpower. However, they have little or nothing to say about Output Willpower.

VI. Conclusions Regarding James’ Theories of Willpower

By this point I hope to have accomplished the goals with which I started in investigating James’ theories of willpower as set forth in *The Principles of Psychology*. First, an assessment of James’ theories of willpower has helped to clarify the specific mental and motivational features of those behavioral performances that are the exercise of Output Willpower. That is, the assessment of JPTW and JSTW showed them to be deficient in crucial respects as an explanation of Output Willpower, and this both helps to clear the way for further advances in the pursuit of an explanation of Output Willpower and provides hints in the direction such an explanation will have to move. Specifically, I have attempted to clarify the ideas that: (i) the operation of Output Willpower does not involve a change in the ideas present in consciousness about the ends of action, (ii) Output Willpower involves some added motivation other than those engaged in the first instance, (iii) within the folk theory, willpower is distinct from a motivational “device,” and (iv) Output Willpower seems to involve something akin to a change in attitude toward the object of the will, such that something like a resolution is required to obtain it.

Second, I have shown how James’ general theory of motivation could lead him to distort his description of willpower. That is, James’ basic motivational theory is that an idea in consciousness automatically triggers the motivation that impels action. Given the
need to fit all phenomena within this general conceptual framework, James is inevitably led to turn a discussion that ought to be about how willpower causes action into a discussion of how willpower causes ideas to dominate the mind, the latter being for James an aspect of how an agent makes a decision to act. Moreover, this background theory also tends to prejudice his discussion of the folk theory and phenomenology, at times making the effort involved seem primarily a contest of ideas, rather than a struggle to take action.

Third, we saw that James was on the road to providing an explanation for a class of cases falling under the heading “exercises of willpower,” just not the class of cases he set out to tackle, the latter being exactly that class of cases that I wish to explain. Nonetheless, an analysis of James’ work helped to distinguish two classes of cases: “Input Willpower” and “Output Willpower.” In making this distinction the target of my explanation becomes clearer and much needless confusion and unnecessary work is avoided.

Fourth, James, and those who have pursued the basic line of explanation offered in JPTW or who otherwise seek to explain Input Willpower, provide examples of how one can attempt non-metaphysical explanations of willpower without crossing over into issues of the metaphysics of free will or volition. I have tried to suggest some ways in which a little more care needs to be taken in this regard and how I hope to avoid these pitfalls. The challenge, going forward, is to see whether there is any prospect for a similar kind of explanation for the more recalcitrant class of cases at which I will aim my explanation, i.e. cases of Output Willpower.
VII. Summing Up The Concept of Output Willpower

I now begin the task of pulling together the features of Output Willpower and elaborating on them a bit. In so doing, I will not attempt to define the necessary and sufficient conditions for the application of this concept. Many things may be called willpower in the course of everyday discussions. Even those that have nothing to do with internal resolve may be called "willpower," if a recent article in a popular magazine on "Simple Secrets to Instant Willpower"\textsuperscript{51} is any evidence of ordinary usage. Hence, there is no obvious reason to think that willpower, or even Output Willpower, is a sharply defined concept with necessary and sufficient conditions for the correct application of its use. Further, even if for reasons having to do with broader philosophical theories, one thinks that any meaningful concept must be in principle a clearly definable term, the abysmal track record of this endeavor in relation to other concepts is sufficient reason to avoid postponing the investigation of the phenomenon until it is precisely defined.

Rather, I will describe key features attributed to Output Willpower. I will try to present these features and describe their relevance to the attribution of willpower to an individual in such a way as to support the following claim. An adequate explanation of attributions of willpower, in the sense of Output Willpower, to agents by other agents requires that each of the specific characteristics or attributes of willpower herein described be accounted for. It may be possible that something can be Output Willpower without exhibiting all of the following features. However, I contend that such cases will be the exception, and that one would have to give a convincing explanation of why, in such cases, one was not either offering a deficient account of Output Willpower or,

---

\textsuperscript{51} Karen Astrid Larson, "Simple Secrets to Instant Willpower," \textit{McCall’s} vol. 126 no. 5 (Feb. 1999): 76-84.
alternatively, simply changing the subject. Of course, as I shall try to make clear, some of these characteristics are more central to the idea of Output Willpower than others.

1. **Overcoming Temptation.** One of the most obvious and commonly noted features of willpower is that it sometimes enables the agent to overcome "temptation." This idea was implicit throughout the discussion of both folk theories of willpower and purportedly scientific theories of willpower. However, a further word should be said about temptation. The idea of overcoming a temptation means, at a minimum, that but for the application of the means with which the temptation was overcome, then, all things being equal, the agent would have chosen a reward dubbed a "temptation." Returning to BMA, it must be the case that an instance of genuine, as opposed to merely apparent, temptation must be one wherein, at a point when two or more competing rewards are available to the agent, the reward value of the tempting option is greater than the reward value of the option perceived as best. Typically this involves a temporary change in preferences that, at sufficient remove before or after the decision point, we believe does not reflect our best judgment about what we have most reason to do. I will have more to say about this in what follows. In sum, one overcomes temptation when, by some means or other, he deliberately takes the course of action incompatible with obtaining the tempting end.

2. **Clear-Eyed Resistance to Temptation.** As the foregoing descriptions should make clear, Output Willpower concerns a kind of strengthening which allows one to overcome the temptation, and thus needn't rely on a weakening or downgrading of the temptation. In cases of willpower, all things being equal, the temptation retains attraction relative to the endorsed reward. Thus, when one overcomes temptation through
willpower, one has neither lost the desire for the temptation, nor lost sight of the temptation in such a way as to weaken its effect on the motivational system. James captured this idea in stating that in the exercise of willpower “both alternatives are steadily held in view, and in the very act of murdering the vanquished possibility the chooser realizes how much in that instant he is making himself lose. . .”\textsuperscript{52} For this reason, Ainslie describes the phenomenon as “clear-eyed” resistance to temptation. If the temptation no longer exercises its pull on the motivational system either through a change in tastes, a dampening of response strength to the reward, or simply through having a different view of the reward token or reward type which makes it unappealing, or having forgotten about the existence of the reward, then no extra “force” of will is needed to overcome it. The temptation has subsided and the motivational challenge to the will has vanished. There is no genuine temptation to overcome.

Suppose, for example, that Jane sat down to dinner full of desire to round it off with a large, delicious piece of chocolate cake, but, unfortunately, eats something during the course of dinner that makes her ill. The thought of eating chocolate cake, or any other food, makes her queasy. When Jane declines John’s offer of a slice of chocolate cake, we are not likely to attribute strength of will to her.

3. \textit{Added Motivation}. Output Willpower involves an added motivation. It is not that the temptation is downgraded, but rather that some additional motivation is utilized which enables the agent to obtain the endorsed end. In attempting to flesh out this aspect of willpower, Elster writes that this boost in motivation “permits you to climb

\footnote{James, \textit{The Principles of Psychology}, Vol. II, 534.}
higher uphill slopes” as opposed to “lower[ing] the height of the slope to be climbed.”

James writes that “our spontaneous way of conceiving the effort … is as an active force adding its strength to that of the motives that prevail.” Whatever the ultimate explanation, the folk concept of Output Willpower, and the apparent experience of Output Willpower, involves some added motivational force.

4. Special Motivation The added motivation operative in output willpower has been described as “some new force distinct from the impulses primarily engaged.”

This is what I will call the “Special Motivation Requirement.” In the first instance, the motive must be one other than those attaching to the primary objects of the decision. But more than this alone is required. Earlier, I stated that a theory explaining willpower that failed to account for this feature of willpower was subject to the Ordinary Motive Objection. That is, such an explanation relies on ordinary motivations, identifiable by the agent as associated with the primary objects of choice or some other motives connected with the choice in some ancillary way. For instance, in deciding whether to turn in a diamond ring that one has found on the floor while shopping, one might be motivated to overcome the temptation to keep it by the desire to preserve one’s reputation for honesty in the eyes of one’s companions. In this case, one is motivated to overcome the temptation by an ordinary motivation—obviously identifiable as something other than willpower even for a fairly unreflective person.

---


There is a second, related objection, based upon yet a stronger interpretation of the idea of a “special motivation.” This position is that all the tokens of such unique motivations operative in instances of the use of Output Willpower are tokens of the same type. In other words, in this view Output Willpower involves a single, unique added motivation. This seems to fall naturally out of the folk concept of willpower in that it seems to see the will as a unique causal force. Any theory of Output Willpower that posits added motivations of different types in different instances is thus subject to what I will call the “*Multiple Motives Objection.*”

Traditionally, the special motivation involved in willpower were explained in terms of the mysterious “volitions” to which Ryle refers, but if this aspect of the folk-concept is to be explained within the framework of BMA, another answer will have to be found. Given BMA, this answer will have to be in terms of rewards and the motivations to which they give rise, or, alternatively, this aspect of the folk-concept will have to be explained away and discounted as the result of an illusion.

5. *Resulting From a Decision.* Another feature of the exercise of Output Willpower is that the added motivation should directly result from a decision to act. Any process of overcoming temptation that is automatic or otherwise appears to the agent to be out of his control should not be regarded as an exercise of will. This is central to the idea of exercising strength, commitment or determination. One triumphs over temptation through inner resolve, rather than being carried along on a fortuitous boost of motivation that happens to support one’s aims. One decides on an end and, at the same time, garners and utilizes the strength to achieve it. The connection between the resolution to take action and the action itself that was highlighted in James’ discussion of JSTW and the
investigations of Ach offers a preliminary suggestion as to how this connection might be
centralized.

We can also say more about how the motivation in Output Willpower must be
connected with the choice. The use of willpower is not experienced as a decision that
trips some motivational automatism that, once engaged, operates like a reflex or a
compulsion. Such would no more be an effort of will than deliberately tripping one’s
own patellar reflex, causing one’s leg to rise, is an intentional action of raising of one’s
own leg. Rather, the agent deliberately “exerts an effort of will.” He deliberately
engages the relevant motivations and identifies with the continuing exertions that result.
The motivation must arise directly from one’s resolve to act. Any theory that involves
the agent using some apparent intervening mechanism, or otherwise indirectly causing
himself to act in a certain way, is subject to the Device Objection.

Putting a finer point on this issue, Output Willpower is not perceived by agents as
involving this kind of indirect self-manipulation, whether by tripping an automatism, or
through some other less dramatic means such as manipulating one’s motivation. The
example of the swimmer purposely overheating himself in order to cause himself to have
sufficient motivation to dive into the pool illustrated this point. This issue of the
connection between Output Willpower and decision I have called the “Linkage Problem.”
In the course of this dissertation I will further develop an account of how best to frame
and how best to solve the Linkage Problem.

I must stress that by “willpower” I do not intend to indicate a concept of
“volition,” such as that criticized by Ryle. As already described, the phenomenon I wish
to explain is not a supposed element of every intentional action. Rather, as it is
commonly understood, Output Willpower is employed selectively, where the circumstances require a special effort to stick to one’s end. Further, as shall be seen in some detail, it is far from the only means at the agent’s disposal to enable him to stick to his endorsed end in the face of temptation. So in speaking of willpower, I am speaking of a specific motivational phenomenon resulting from intentional choice to engage the relevant motivations.

In making these claims, I offer no account of what it means for an action to result from choice, but rely on ordinary attributions of freedom of choice to clarify the notion of willpower. It may, in fact, be that there is no place for free choice in any robust sense under a causal theory of motivation such as BMA. Regardless of the final answer to such questions, the fact remains that the exercise of will is commonly understood as flowing from choice, rather than merely being a surge of motivation that operates upon or through an agent regardless of his choice or decision.

6. **Internal means of Resisting Temptation.** The exercise of willpower must be in some sense internal. Ainslie has described this sort of resistance to temptation by willpower as “wholly mental impulse avoidance.”\[56\] I think a better way to put this is that Output Willpower involves an added motivational force, and that it is this force that is the immediate cause of the action taken in choosing the endorsed, rather than the tempting end. At the same time, this force is set in motion directly by the agent’s choice. Hence, Output Willpower is a wholly “internal” process in that it is operative through these mental and motivational events and/or processes. This can be made clearer by way of

---

\[56\] Ainslie, *Picoeconomics*, at 154.
example. In the first instance I will provide an example of what willpower is not, and then move on to provide a positive example of willpower.

Consider the case of a woman, Jane, who is trying maintain a strict diet for one month, but has a definite weakness for chocolate cake. Suppose Jane now has no desire for chocolate cake. She may have just eaten a very large slice. However, she realizes that, over the next few days, her appetite for cake will return. Wishing to avoid succumbing to the temptation, she makes her husband, John, promise to prevent her from eating chocolate cake for the next month. John may throw away all chocolate cake he finds in the house, control Jane's spending, and monitor her dining and her trips to restaurants or grocery stores. As her appetite for cake, and her irritation with John's interference, increases, Jane pleads with John to forget her "foolish" request. She tells him he should stop interfering and that he should be reasonable. Jane tells John he need not stick to his promise. However, John refuses all her arguments and entreaties and continues with his best efforts to live up to his promise despite Jane's protestations. Even if John is successful in preventing Jane from eating cake for one month, few would credit Jane with exercising willpower to resist the temptation of the chocolate cake. Rather, if anyone has exercised resolve or strength of will it is John.

The "force" with which the temptation has been resisted has come from outside Jane. To be an exercise of will, this resistance must, in some sense, come from "within" Jane. Her resistance must arise directly from her "inner strength" and not be effected or mediated by means by external forces not under her direct control. Thus, in any case where a causal force that Jane, or another person, sets up to operate outside of Jane's own motivational system and that determines her action so as to prevent her from succumbing
to temptation her behavior would not likely be called an example of Output Willpower. Further examples illustrating and confirming this point can easily be constructed. For instance, if Jane enforced her diet by having her jaw wired shut, her subsequent abstention from cake would not be credited as an exercise of willpower. For it is the restraining device that does the work, not Jane’s resolve. Of course, this is not to say that Jane may not have exercised some willpower in having the device attached, but that is another matter.

Had Jane, on the other hand faced up to the temptation of the cake directly, she might have been credited with an exercise of Output Willpower. For example, Jane might have seen cake offered on a menu, or on sale in the grocery store, and yet resisted it without the aid of her husband or other external restraining forces. Though she wanted it, she may have simply resolved not to buy it. In such a case, she might be regarded as having demonstrated willpower. The point is that in exercising willpower, Jane’s action in contravention of temptation must be the direct result of the input from her motivational system, not the causal result of some force or restraint outside her own motivational system.

7. **Willpower as a Causal Power.** The argument to a causal account is a short one. First, the exercise of Output Willpower has been described as an instance where, but for the exercise of this power, one would have succumbed to temptation. This means that the power of the will (whatever the “will” may be) is a causal power. It is part of the notion of will that it be causally effective in one’s behavior in overcoming temptation. People tell each other and themselves that they should overcome obstacles, kick bad habits or otherwise achieve their aims by the injunction: “use your willpower.”
The idea here is that there is some force or capacity which the agent has, which, if applied, would allow him to prevail in the face of temptation. Similarly, after having obtained a certain kind of success over temptation, an agent will often explain this success in terms of willpower. These ordinary facts of experience are simply unaccounted for by Ryle, and his account of willpower seems to unjustifiably omit this common aspect of experience. It is this fact, the fact that willpower is commonly conceived of as a causal force, that makes it a candidate for a causal/scientific explanation, and suggests that there is an inherent flaw in the concept if it cannot be so explained.

I have said that the exercise of willpower results from decision. This seems to suggest a limit to causal explanation. While the will may be causally effective, it is set in motion by a decision, a free decision, which resists causal explanation. One is faced with the old conundrum of the free will/determinism debate. That is, once we have started down the line of causal explanation we are forced into a regress. If we are not to bring the explanation to an arbitrary halt, we must ask what caused the exercise of will, or, alternatively, what causal forces are at work in the exercise of will. This regress appears to threaten the notion of a voluntary decision to act in general, and, therefore, the notion of willpower in particular.

I will not take up the long-standing debate between those who hold that some concept of freedom of will or moral responsibility is compatible with determinism and those who hold that it is not. It is simply too large a topic and would sidetrack the investigation into the phenomenon of willpower that is the topic of this dissertation. In

---

searching for a causal explanation of willpower, I will leave the notion of "decision" unanalyzed in causal terms. I will simply rely upon our ordinary criteria for deciding when to attribute a decision to an agent and when not to. Whether such attributions can ultimately be justified, on a causal account, is a topic that cannot be entered into here. This limitation in scope seems particularly fair in this context since, as we shall see, all the competing candidates for an explanation of willpower do the same. Therefore, without offering an analysis of the concept of "decision," I will simply proceed on the basis that whatever the notion of "willpower" may designate, if it effectively designates at all, the answer to the question of the efficacy of this power must be a causal one. In other words, if the folk concept of willpower is going to be explicated in a coherent way, then it must be by means of a causal explanation.

8. *Force of Willpower a Matter of Degree.* It is also a part of the folk concept of willpower that whatever sort of motivation it is, it may be applied to greater or lesser degrees. As James noted in setting out the folk theory, this motivation "may be of various amounts" and is "adventitious and indeterminate in advance."\(^{58}\) Willpower is a scalar phenomenon. It is not merely a matter of throwing a switch on or off which applies a fixed quantity of motivation. A person can strive more or less, exert his will more or less. Having failed at an effort once, he may return to the same task with a greater effort of will and succeed where before he had failed. Similarly, the effort of will is typically experienced as proportionate to the task. Breaking away from the TV during a very interesting TV program to listen to what your spouse is saying presumably takes less an effort of willpower than hurling yourself over the top and into enemy gunfire

---

during trench warfare. Hence, the scalarity of willpower is another formal feature of the concept that must be accounted for.

9. **Not Rationally and Ethically Mandated.** I have already described the idea of willpower as involving the overcoming of temptation, but more must be said as to what is meant by “temptation” and the connection, or lack thereof, between willpower and the agent’s judgments about ethics and rational prudence. At first, it seems built into the concept of “temptation” that the tempting reward is that which one judges one ought not to indulge in. In other words, given a set of alternative rewarding events, one (at the least) judges the tempting alternative not to be not the best of the available options according to whatever normative criteria one is applying to the decision. Typically this is the case. However, in the context of the analysis of Output Willpower, it is not necessary that “temptation” always be understood in that way. For though it may be true that willpower is often employed to enable the agent to act in accord with his all-things-considered judgments about which course of action is ethically best or most in accord with rational prudence, I contend that this is not a necessary feature of Output Willpower. “Temptation” is best thought of in this context as a temptation to “weakness.” Here, I use the term “weakness” primarily because there is no convenient alternative. I mean only to use a term applicable across a broad range of contexts as described below.

Consider, for instance, one who acts from spite. Suppose a man, call him Jim, proposes a plan at a meeting to solve the current financial crises faced by his employer. Jim’s colleague Roger bluntly states that Jim’s plan is foolish and goes on to convince Jim’s boss and colleagues that it will never work. Humiliated, Jim may plot and plan and lie in wait for months or years for his chance to gain revenge by humiliating Roger.
Carrying out the plan of revenge may require a determined and sustained effort of willpower, but it may also be to Jim's overall detriment.

It may be rationally prudent, all things considered, for Jim to move on to other matters. Jim might know that he would be better off if he just "let it go" and "chose his battles" more prudently. Similarly, in a calm discussion, Jim may well agree that revenge is not an ethical course of action. He may agree that the best thing to do from an ethical standpoint would be to speak to Roger directly in an attempt to both express his dissatisfaction with how he was treated and to attempt to put his work relationship with Roger on a better footing. It is a commonplace that determined efforts of will needn't conform to all-things-considered judgments of ethics or rational prudence. In fact, one who sets out on such a course but later yields to the nagging of conscience or the voice of rational self-concern may well think himself weak, or be thought weak by others if they are engaged in a common plan. In this case, the "temptation" is not to act against self-interest or morality, but rather to act in accord with it. The "temptation" is the temptation to weakness, not the temptation to evil. The idea here is that "weakness" is a kind of lack of resolve, most easily characterized as an unwarranted failure to do in the future what one currently intends to do.

There are probably several inter-related but distinct, sometimes conflicting, criteria as to what constitutes such weakness. However, some things are clear about how

59 There is an urge to say that this cannot be the case despite the many easily imagined examples that justify this view. I believe that this urge arises from the more general presumptions that agents cannot deliberately pursue that which they believe to be wrong, or, alternatively, that agents cannot deliberately do that which they believe to be rationally imprudent. I believe both of these presumptions are belied by common experience, but an all-out attack upon them is beyond the scope of this dissertation. For criticism of these presumptions, see Michael Stocker, "Desiring the Bad: An Essay in Moral Psychology," Journal of Philosophy, vol. 76 no. 12 (Dec. 1979): 738-53. See also Henry Alker, "Will Power," Analysis, vol. 21 no. 4 (March 1961): 78-81, on the moral neutrality of willpower.
the concept is employed. Such judgments of weakness may be consistent or inconsistent with prudential or moral judgments about what behavior is best, as shown by the example of Jim and Roger. Similarly, one may use willpower in the service of short-term or longer-term preferences, such that one is seeking to overcome contrary motivations that may be either relatively short-term or long-term. For instance, Jim’s interest in revenge may have been short-term relative to his competing, longer-term ends.

Of course, this doesn’t fully explain the idea of acting from “weakness.” At its basis is the notion that at a given point in time the agent resolves to pursue and obtain some end, but then fails to obtain that end. The failure is explained in terms of the agent being insufficiently motivated to obtain the end, and not, for instance, due to forces outside himself and thus beyond his control. If Jim failed to carry out his plan of revenge because Roger unexpectedly left the company before it was possible to carry out his plan, Jim would not think of himself through having failed through weakness, i.e. through lack of motivation or effort. The failure at issue is one of motivation. Given what I have already stated about the lack of necessary connection between willpower and ethical and prudential judgments, it is clear that the weakness at issue isn’t that of akrasia, which is typically analyzed as a failure to act in accord with one’s ethical or prudential judgments,\textsuperscript{60} or, on some theories, as acting in accord with short-term preferences at the expense of longer-term preferences.\textsuperscript{61} What kind of failure of action is it then?


\textsuperscript{61} Elster, Ulysses and the Sirens, Sec. II.5; Ainslie, Picoeconomics, at 154.
It may be that the relevant notion of weakness is specifiable in terms of some complex criteria of application drawn from ordinary language. However, I will not endeavor to so specify the criteria for the concept here. For my purposes, "weakness" may be understood in purely formal terms. That is, there is some end, x, which the agent resolves to obtain. If he lacks motivation to obtain that end, then his motivation or "will" can be said to be weak in that regard. Willpower in some way augments the agent’s motivation to obtain that end, and, ideally, enables him to obtain it. For purposes of discussion, I will call the end that the agent resolves to obtain the relevant point in time the "endorsed" end or aim or reward, and the alternative course of action (or inaction) which may lure him away from the successful pursuit of that end as the "disfavored" or "specious" end or aim or reward. Defined in this way, the terms are, of course, time and context relative. However, I have sufficiently specified them to allow the analysis to proceed.

VIII. Summing Up the Chapter and Moving Forward

We now have a set of criteria for Output Willpower, the kind of willpower that I intend to explain. Given these criteria, in going forward to assess some of the various explanations offered for this phenomenon, I will focus my efforts on theories that involve ways of overcoming temptation where the operative factor is an added motivation. However, I will show that the competing theories that this restriction admits to the field fall to ground on the more complex criteria of Output Willpower that I have called the resolution of the "Linkage Problem" and the "Special Motivation Requirement." I will also offer and assess additional objections to these theories that, while not directly related to the initial criteria of Output Willpower, go to the very heart of the functioning of these
mechanisms. In the end, I will offer a theory that meets all of the requirements of Output Willpower that I have set out.
Chapter Three
Personal Rules and Private Side Bets

I. Introduction

There has been little written in the way of a scientific account of Output Willpower.\(^1\) Perhaps unique in providing a detailed theory is the work of George Ainslie. Ainslie has created a “model of willpower” that “creates a very specific hypothesis of ‘strength.’”\(^2\) Ainslie describes the mechanism involved as being or utilizing “personal rules” in a specially defined sense. I will argue that while Personal Rules\(^3\) are interesting and important, it is a relatively poor explanation of Output Willpower as I have described it. Of specific interest is the connection of Ainslie’s ideas with the Linkage Problem and the progress that can be made by taking Ainslie’s theory, not as a final explanation of Output Willpower, but as a launching point for further exploration.

II. The Basic Personal Rules Mechanism

The Personal Rules mechanism may be best explained by means of an illustration offered by Ainslie. In the example, the choice to be made by the agent is between staying up two hours later in the evening to watch TV, or getting a good night’s sleep and feeling well-rested at work the following day.\(^4\) Initially, the agent may most want to go to bed

---


\(^3\) For purposes of showing that this is a defined term, and for consistency, I capitalize “Personal Rules” although Ainslie does not do so.

on time and be well rested. However, as evening wears on and bedtime approaches, the agent may most want to stay up and watch TV, and thus succumb to the temptation to do so. This is represented in figure 3.1(a)\textsuperscript{5} by the hyperbolic curves. The \textit{x}-axis corresponds to the time delay until the agent receives the reward. The \textit{y}-axis indicates reward value. The reward value to an agent of a given reward at any particular moment is represented by a point on its respective curve. The curve that is higher at any given moment represents the agent’s dominant motivation or preference. Since the dashed curve representing the motivation to stay up and watch TV crosses over and rises above the solid line curve representing the motivation to get a good night’s sleep, the agent’s preference will switch to watching TV over going to bed as the evening wears on. Since he will be presented with the option of staying awake to watch TV when its reward value is dominant, that is the reward he will ultimately choose.

In order to understand these graphs, it is helpful to review the background theory that they presuppose. Ainslie has argued extensively for the hypothesis that for all animals the nearer in time some future rewarding event is, the stronger will be the behavioral response of the animal in seeking it. Conversely, the further in the future a rewarding event is, the less the animal will exert itself to obtain it. Thus, there is a “time discounting of the reward value” of the future event. In fact, it appears that future rewards may be discounted at a very severe rate, more specifically, a hyperbolic rate, if not ameliorated by some process or mechanism, or so Ainslie argues. Hence, the phenomenon is referred to as “hyperbolic time discounting.” It is because of hyperbolic time discounting that the reward curves cross. While I will not assume Ainslie is correct

\textsuperscript{5} Figure taken from Ainslie, and Haslam, “Self-Control.”
in this regard when discussing the views of other theorists or in giving my own account of willpower, it is useful to keep hyperbolic time discounting in mind when considering Ainslie’s theory of Output Willpower (even though we could reconstruct a version of Ainslie’s theory of the mechanism of Output Willpower without relying upon the discounting hypothesis).

Note that the choice that would maximize the agent’s reward as it is measured "objectively" would be for him to go to bed early. An “objective” measure of a given reward is what Ainslie describes as reward value at a given temporal delay, where that delay is held constant. That is, the reward value of each temporal segment of a given reward could be assessed as if each such segment were to occur at the same temporal distance from the agent. To compare two rewards, the reward value for each would have to be determined at the same time delay.

The objective value of the rewards in the example is represented graphically in Figure 3.1. We assume that each temporal segment of a given reward would be valued equally by the agent if they occurred at the same time delay. Thus, the value of a reward at a given time delay could be represented by a horizontal line drawn on the graph. The distance of the line from the x-axis denotes the reward value at a given time delay, and the length of the line represents the actual duration of the rewarding event. A column is formed by drawing two lines rising from the x-axis and intersecting the ends of the horizontal line. The area within this column represents the objective quantity of reward. In this example, the objective quantity of reward per hour that the agent would get for the two rewards is equal. In objective terms, the reward for going to sleep on time is greater than the reward for staying awake and watching TV, because the time span from the time
the agent gets up for work (7:00 a.m.) to the time he returns from work (5:00 p.m.) is
greater than the two-hour time period during which he would watch TV. Thus, there will
be a net loss in objective reward value when the agent succumbs to the temptation to stay
awake and watch TV.

Suppose the agent realizes that he will face this same decision, regarding whether
to go to bed early or not, on a nightly basis. On each occasion he faces the possibility of
succumbing to temptation and losing the larger, later reward. In such situations, the
agent lacks full information about his ability to resist the temptation of the smaller reward
in future choices. It may be difficult for him to predict how he will behave.

One salient piece of information that the agent will soon have at his disposal is
whether he successfully resists the temptation in the present choice. Whether he is able
to resist temptation in this instance may thus be regarded as a predictor of his future
success in overcoming relevantly similar temptations. If he succumbs to the temptation
of the smaller reward in the present instance, his expectation that he will be capable of
resisting temptation in relevantly similar choice situations in the future decreases.
Similarly, if he resists the present temptation, his expectations that he will resist
relevantly similar temptations in the future, and gain the larger, later reward involved in
those choices, increases.

Once the present choice of the agent takes on this significance as a test case for
future success, more is at stake for him in the choice. If he succumbs to the temptation to
stay awake for an extra two hours, he pays an immediate cost because of the decrease in
his expectations regarding, not only his ability to go to work well-rested on the following
morning, but also in his being able to go to work well-rested on an indefinite number of
future mornings. Similarly, if he resists the temptation to stay up later, he will, in the present, increase his expectations of enjoying future days at work feeling well-rested and fit. These losses or gains accrue to the individual at the time of the initial choice in the form of diminished or increased expected rewards. As in standard utility theory, increased expectation that a reward will be received increases the reward’s value. Decreased expectation that a reward will be received decreases its value. Thus these future rewards, i.e. the rewards of going to work well-rested on several or many days in the future, represent an additional benefit perceived to be at stake in the decision.

The “Personal Rule” is, in effect, “In all cases of type A (e.g. cases where I must choose between staying up and watching TV two hours extra and going to bed early for a good night’s rest) I will (or must) always do x (e.g. go to bed early) rather than y (stay up late to watch TV).” The thought that harnesses the motivation which gives the rule its force is: “If I fail here, I will probably fail in all future choices of type A, and lose the greater rewards involved in choosing x over y for the series of choices between x and y, but, if I succeed here, I will probably succeed in choosing x over y in choices of type A in the future.” In other words, once the agent takes this current choice to be predictor of future choices, he will feel that what is at stake is not merely, e.g., proper rest on this night, but the expectation of many nights’ proper rest, as opposed to many nights of staying up and watching TV.

---

6 Ainslie is vague about what form the Personal Rule actually takes. At one point he writes that “if a person wanted to commit himself to keeping his shoes shined, he might make a rule that he would shine them every morning before breakfast.” George Ainslie, “Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control,” Psychological Bulletin vol. 82, no. 4, (July 1975): 463-94, 481 (emphasis added). Ainslie’s wording is ambiguous as to whether the man is making a prediction or making some kind of resolution. The importance of this distinction is a topic I will pursue in what follows. For now, I simply note the problem. Typically, Ainslie simply operates in terms of what I referred to as “the thought that harnesses the motivation,” and appears to either be treating this as the rule or he simply does not specify the rule itself.
Ainslie represents this process of aggregating rewards for a given choice as a simple summation function. This effect is depicted on the graph, Figure 3.1(b). In comparing these two aggregated rewards, the individual's desire to get decent rest on this series of nights will be so much greater than his desire to stay awake and watch TV on this series of nights, that their respective reward valuation curves will never cross. This indicates that the agent's preferences will never switch as he approaches the time at which he must turn off the TV and go to bed. The agent will not succumb to the temptation to stay awake in the current choice, and will enjoy the larger reward of going to work well-rested on the following day. In general we can say that because of this device the agent will be motivated to overcome a variety of short-term temptations since he now has more at stake in resisting them. By seeing a series of future rewards as being at stake in a given choice situation, and staking one's expectation of receiving them on one's ability to resist temptation in the present situation, one is able to resist that temptation because the potential gains or losses represented by the increased or decreased expectations, motivate him to do so.

The predictions embodied in these Personal Rules are self-fulfilling. The change in an agent's expectations, in altering his motivational states, materially changes his prospects for realizing future rewards. This follows from the fact that when expectations of obtaining a rewarding event are raised or lowered it affects the current reward value of the rewarding event to the agent. As previously noted, one factor affecting the extent to which a reward is discounted is the agent's estimation of the probability of its occurrence. The shift in reward value corresponds to a shift in the agent's motivation to seek the

---

rewarding event. Therefore, if an agent believes he cannot obtain a rewarding event, he is less likely to obtain it because his motivation to seek it is diminished. This loss in motivation may lead the agent to seek other rewards that were specious vis-à-vis the previously undiscounted, endorsed reward, but that now exert a greater motivational influence. The contrary would be true were the agent's expectations of attaining future endorsed rewards to increase. The increase in the agent's expectation of obtaining the endorsed reward would actually increase the likelihood of obtaining it. Thus, future rewards, and not merely the expectation of them, turn out to be at stake when the Personal Rules mechanism is engaged because the agent's diminished or increased expectations of a reward tend to be self-fulfilling.

Thus far, nothing has been offered by Ainslie's theory that serves as an explanation of willpower. The reason for this is simple. The mechanism behind Personal Rules as described thus far is determined by chance, rather than choice. The motivational power which these Personal Rules supply is put into play by the motivational pull of a set of reward contingencies which the agent finds himself subject to, as opposed to his having established them by his own choice or design. Rather than calling upon some inner resolve, he is merely reacting to a "discovery"8 that a given decision is likely to be a predictor of future decisions. It is the structure of the decision points and the agent's awareness of them that determines his motivation, and in this respect he is merely passive. An explanation of Output Willpower requires more.

---

8 Since belief that a given decision is a predictor of future decisions will, on this model, tend to make it a more reliable predictor, the agent does not merely discover the rule, but also helps to give it force by his belief in it. Thus, since the Personal Rule operates on the basis of a kind of self-fulfilling prophecy, one cannot really say that it is simply discovered as an objective fact existing wholly independently of the agent's beliefs.
III. Some Varieties of Personal Rules Mechanisms

A. Overview

So far I have described what I will call "Predictive Personal Rules." Predictive Personal Rules are Personal Rules premised on a belief in the value of a given decision in predicting the outcome in future situations of decision. Ainslie suggests that an agent may come to employ a Predictive Personal Rule in an intentional manner by focusing his attention on choices that are likely to be interpreted by the agent himself as a bellwether of future behavior. 9 This introduces the element of choice into the use of Predictive Personal Rules, thus apparently countering my initial objection to Personal Rules. In addition, however, Ainslie's account also envisions a distinctly different form of Personal Rules that may be intentionally deployed. According to Ainslie, the agent may formulate his own Personal Rules in order to intentionally boost his own motivational capacities. It is these strategically created Personal Rules that I refer to as "Prescriptive Personal Rules."

Ainslie writes:

... If the organism detects the similarity of a series of choice points, its behavior at one point can become a cue in predicting behavior at other points ... Insofar as choice provides cues about future choices, it will be affected by the rewards that depend on these future choices. ...

This hypothesis suggests two routes by which an individual can acquire a private skill for controlling impulses, without relying on other people or the mechanical properties of the environment. [1] He can learn to attend to cues in his behavior that predict future behavior, thus finding secondary reward to bet against impulses. [2] But in cases in which this does not provide enough reward, he can also create categories of gratification-delays behavior whose members will stand or fall together. That is, he

can arbitrarily propose a bet which, if it has the necessary properties, can motivate his future behavior just as well as the bets set up by chance.\textsuperscript{10}

The first alternative describes how an agent may gain partial control over when the Predictive Personal Rules mechanism will be engaged through control of attention.\textsuperscript{11} However, this does not alter the basic operation of the mechanism once such a predictor-choice is attended to. A given instance of the mechanism’s motivational power is still dependent on the agent’s belief in the truth of the motivating thought behind the Personal Rule that governs its functioning. The second alternative, Prescriptive Personal Rules, represents a radical departure from the account of Personal Rules thus far given, and, I will argue, is deeply confused. I will address the failures of each of these alternatives as an explanation of Output Willpower in turn.

\textbf{B. Problems With Intentionally Employed Predictive Personal Rules}

Predictive Personal Rules engaged with the aid of control of attention suffers from some of the same shortcomings that control of attention does generally as an explanation of Output Willpower. First, as already discussed, control of attention may be subject to the Device Objection. Since Predictive Personal Rules are a special mechanism, not a part of all action like James’ hypothesized motivating ideas of action, they are particularly vulnerable to this objection. In cases of attending to potential predictor-choices, one is selectively attending to such situations of choice. In so doing, one is attending to that which moves him willy-nilly. He makes the choice to attend to the stimulus, he does not choose whether or not to be motivated by it. This seems to be a

\textsuperscript{10} Ibid. at 479-89 (emphasis added).

\textsuperscript{11} See Chapter Two.
case of creatively manipulating one’s motivational system, rather than choosing to take an action, or exerting oneself to achieve an end.

Here the resolution to take action (if there is one) either has to do with the control of attention itself or comes after the Personal Rule is discovered/constructed. In the former case, the resolution is misplaced vis-à-vis any explanation of willpower that hopes to go beyond the account offered by control of attention. In the latter case, it does not properly address the Linkage Problem. What the latter scenario amounts to is simply the agent first realizing that more is at stake in the decision about which action to take than he initially realized (i.e. he has the realization that there are the additional stakes involved in the choice as framed by the Predictive Personal Rule) and then taking the endorsed course of action in light of those added stakes. There is no sense in which a resolve of some kind does any work in the formation or operation of the Personal Rule. The agent simply does what he is most motivated to do in light of his new grasp of the outcomes at stake. Positing some kind of personal resolve to take action in this case—beyond being naturally inclined to move in the direction of the greater reward contingency after the formation of the Personal Rule—is merely *ad hoc* unless some separate explanation of its role is devised.

A second objection to this analysis of Output Willpower is that it fails to explain how willpower appears to introduce into the agent’s motivational system “some new force distinct from the impulses primarily engaged.” In other words, it fails to meet the Special Motivation Requirement. Although new reward tokens are at stake, by definition Personal Rules involve repeated iterations of the same type of decision-making scenario,

---

and this means more tokens of the same type of specious reward being balanced against
more tokens of the same type of endorsed reward. Although the consequences of the
choice is perceived as more grave when the Personal Rule is engaged, this is because
“more of the same” is perceived as being at stake. Therefore, this version of the Personal
Rules mechanism is subject to the Ordinary Motivation Objection. Since different
Personal Rules will involve different linked series of reward tokens of differing types, the
mechanism is also subject to the Multiple Motivation Objection.

There is another class of objections that moves beyond these first three objections
which were based upon the descriptive criteria of Output Willpower. The first of these
new objections concerns the fact that the two series of reward contingencies must be
perceived as endless, without a definite termination point, or, at least, very long. Using
the example of a dieter, Ainslie sets out the problem in this way:

As [the agent] nears the end of a finite series, he will have decreasing
amounts of future rewards to bet against his impulse to eat, until at some
point he will prefer to lose the bet. If he knows that he will not avoid the
last temptation, he will not be able to avoid the one before it; and if he
recognizes this, he will not be able to avoid the next earlier one, and so on.
Depending on his perceptions, this process may or may not proceed like a
row of falling dominoes to make the whole [Personal Rule] worthless.\(^{13}\)

Thus, if the series of choice points falling under a given Personal Rule is finite, and the
agent knows this, the Personal Rule may unravel.

Intuitively, this kind of unraveling is more likely where the series is relatively
short, and may not be so likely where the series is quite long. For it seems that the agent
would not lose the efficacy of the entire Personal Rule merely because at some distant
point in time, with one or only a few choice points left, the Personal Rule would lack

\(^{13}\) Ainslie, “Specious Reward,” at 481. See also Ainslie, Picoeconomics at 162.
efficacy because the endorsed rewards bound together no longer form a sufficient motivation to avoid the temptation. However, a meaningful conclusion in this regard would require empirical studies beyond the scope of this dissertation.¹⁴ I merely intend to note that what we may call the “Unraveling Problem” is a point of concern acknowledged by Ainslie.

Ainslie proposes a solution to this problem in terms of the agent encompassing choices in more general rules that bind more decision points of various kinds together than could be bound together under the original rule.¹⁵ For instance, the bet about shining one’s shoes every morning could be subsumed under a more general bet about one’s ability to properly dress and groom oneself before going into a public setting. Ainslie is rather sketchy about how exactly this works, but it seems to involve Prescriptive Personal Rules. Given the arguments I will make below regarding Prescriptive Personal Rules, Ainslie’s proposed solution to the Unraveling Problem is in

¹⁴ This question of whether a Personal Rule would unravel in this way, and whether the decisions that constitute the unraveling are rational, even over a long, but perceptibly finite, series, is analogous to the question of whether it is rational for agents to defect, rather than cooperate, in an iterated prisoner’s dilemma where the players know the series of iterations to be finite. That is, it is generally held that it is both rational and empirically likely that tacit cooperation will evolve between players in an iterated prisoners because each player has the opportunity to reward or punish the other player in the next round for his cooperation or defection respectively. Robert Axelrod, The Evolution of Cooperation, (New York: Basic Books, Inc., 1984). However, where the iterations are finite, and where the players know that there will be no opportunity for retaliation in the last iteration, one player would defect at the expense of the other. But then both players know this and each can expect the other to defect in the last iteration. On the same reasoning, there would be no reason to cooperate in the penultimate iteration, since the defection of the other player in the last round is a forgone conclusion. In this way the pattern of cooperation is subject to this backward unraveling in the same way as Ainslie’s Personal Rule. The questions of how agents would actually behave and how they should behave if they are rational in an iterated, but finite, prisoner’s dilemma is an interesting question, but one that I will not address. See Russell Hardin, Collective Action, (Baltimore: Johns Hopkins University Press, 1982), 145-50. For an example of the analogy of prisoner’s dilemma to an intrapsychic case not involving Personal Rules, see Robert H. Frank, “The Role of Moral Sentiments in the Theory of Intertemporal Choice,” in Choice Over Time, eds. George Loewenstein and Jon Elster, (New York: Russell Sage Foundation, 1992), 265-284 For analysis of Ainslie’s comparison of Personal Rules to iterated prisoner’s dilemma see Appendix A.

¹⁵ Ainslie, “Specious Reward,” at 481.
serious doubt. Regardless, we can at least say that if Ainslie is correct in his analysis, the series of decisions points required for a Predictive Personal Rule to work must be infinite, without a clear termination point, or at least very long, whether this is achieved through several increasingly general, linked Personal Rules or through a single Personal Rule.

This makes it look as if Predictive Personal Rules could operate only in relatively limited circumstances involving these lengthy iterations. It seems that in many instances the exercise of Output Willpower does not involve this kind of iterated series of choice points. The case of the hypothetical swimmer diving into the cold water discussed in Chapter Two is an example of the kind of case that supports this point. It seems natural to think that our hypothetical swimmer need not be thinking of future choices about diving into the pool when he forms his resolve to dive in. It seems like a one-time decision. Other cases are even more definite in this regard. Many heroic acts, for instance, are believed by the would-be hero to be one-time events in his life, or at least the product of situations unlikely to recur in the future—either because they result from highly unusual circumstances such as war, natural disaster, criminal violence, or some other unusual circumstance, or because he expects to die in the act of heroism. Therefore, it cannot be an expected sequence of relevantly similar rewards which motivates most classically heroic acts said to be motivated by willpower. Hence, this objection, which I will call the “One-Shot Objection,” shows that for at least some instances of the exercise of Output Willpower Personal Rules cannot be the explanation because such cases involve a one-time decision, not a decision that figures as a part of an iterated series of choice points.\textsuperscript{16}

\textsuperscript{16} Actually, there could be a separate objection, based on cases where there is an apparently finite number of iterated decision points (or alternatively, involves a small number of iterations, but more than one, and
In sum, Predictive Personal Rules are an unpromising explanation of Output Willpower. They fail to solve the Linkage Problem and are subject to the Device Objection. Further, they are also subject to the Ordinary Motive Objection and the Multiple Motives Objection. Finally, they suffer from the Unraveling Problem and invite the One-Shot Objection. Still, to see if a Personal Rules theory can get a better handle on the Linkage Problem, it is necessary to look at Prescriptive Personal Rules.

C. Prescriptive Personal Rules

1. Overview

Prescriptive Personal Rules motivate action in the same way as Predictive Personal Rules: by aggregating the motivational effect of relevantly similar future rewards. The only difference is that one designs one’s own rule rather than simply stumbling into the "realization" that one is facing a predictor-choice or attempting to direct one’s attention toward or away from such sequences of decision points which are otherwise not of one’s own creation. Thus, for both Prescriptive Personal Rules and Predictive Personal Rules, the motivational boost is achieved by aggregating the reward value of several relevantly similar rewards. Therefore, Prescriptive Personal Rules are no less subject to the Ordinary Motive Objection, the Multiple Motives Objection, the Unraveling Problem and the One Shot Objection than were Predictive Personal Rules.

However, the hope is that Prescriptive Personal Rules will make some headway in dealing with the Linkage Problem in general and the Device Objection in particular. This is because Prescriptive Personal Rules are purportedly engaged by choice, and formulated where Output Willpower is operative. However, there seems little profit in carrying this technical point through the dissertation when the One Shot Objection effectively conveys this general kind criticism of Personal Rules.
according to one’s own design. Thus, they have less the appearance of an automatism that one chooses to engage than do Predictive Personal Rules. Unfortunately, Prescriptive Personal Rules—at least as they must be articulated within Ainslie’s Theory—suffer from a more grave defect than any so far listed, indeed, the ultimate defect: they are an impossibility. In what follows, I will argue that Ainslie’s argument for the existence of Prescriptive Personal Rules is simply incoherent. Further, even if per impossible they could function, they would still be subject to the Device Objection.

2. The Impossibility of Prescriptive Personal Rules

A key problem in explaining Personal Rules is that Ainslie gives little guidance as to what such a rule would look like. This problem is particularly acute with Prescriptive Personal Rules. However, Ainslie does seem to offer some suggestion: “For instance, if a person wanted to commit himself to keeping his shoes shined, he might make a rule that he will shine them every morning before breakfast.”17 Presumably the prescriptive rule is: “I will shine my shoes every morning before breakfast.” If this is the rule, then we must next inquire as to what I have already called the “thought which harnesses the motivation which gives the rule its force.”18 For a Predictive Personal Rule, this thought may be captured as: “If I fail here, I will probably fail in all future choices of type A (e.g. choices to shine my shoes before breakfast or not), and lose the greater rewards involved in choosing x over y for the series of choices between x and y (e.g. having shined shoes on most or all days in the future as opposed to the lesser reward of sparing myself a little effort each day), but if I succeed here, I will probably succeed in choosing x over y in


18 See infra. Section II.
choices of type A in the future.” But here, the thought which serves to harness motivation is not predictive, but rather “arbitrarily propose[d].” The question then is: What does this thought that harnesses motivation look like? Ainslie does not tell us. In fact, Ainslie never clearly distinguishes between the Personal Rule regarding what the agent will do and the terms of the relevant motivational thought, nor does he set out the precise relationship between them. Rather than speculate on how we might construct a model of such an arbitrarily proposed Personal Rule on Ainslie’s behalf, I will instead consider whether the idea itself makes any sense.

Ainslie’s explanation is couched in terms of what he calls “Private Side Bets.”¹⁹ A Private Side Bet is the mechanism giving the Personal Rule its motivational force. Ainslie uses the analogy to other forms of “Side Bets” to explain the operation of the mechanism involved. Perhaps the best way to proceed is by drawing out the analogy between Private Side Bets and the more well-known idea of a “Public Side Bet.” A Public Side Bet is “an irrevocable change in the payoff structure” which increases the probability that an agent will carry out a given course of action.²⁰

The idea of a Side Bet first gained currency through the work of Schelling, who provides the following example:

When one wishes to persuade someone that he would not pay more than $16,000 for a house that is really worth $20,000 to him, what can he do to take advantage of the usually superior credibility of the truth over a false assertion? Answer: make it true. How can a buyer make it true? …

[S]uppose the buyer could make an irrevocable and enforceable bet with some third party, duly recorded and certified, according to which he would

---

¹⁹ Ainslie does not capitalize the term. I do in order make it clear that it is a defined technical term.

pay for the house no more than $16,000, or forfeit $5,000. The seller has
lost; the buyer need simply present the truth. Unless the seller is enraged
and withholds the house in sheer spite, the situation has been rigged
against him; the “objective” situation—the buyer’s true incentive—has
been voluntarily, conspicuously, and irreversibly changed. The seller can
take it or leave it. This example demonstrates that if the buyer can accept
an irrevocable commitment, in a way that is unambiguously visible to the
seller, he can squeeze the range of indeterminacy down to the point most
favorable to him.21

As is often the case with Schelling’s work, there is a great deal packed into this
brief example. The first point to note is that Schelling is primarily interested the fact that
such a strategy of commitment would allow the agent to achieve a superior bargaining
position vis-à-vis another agent. In contrast, I am interested in the explanatorily prior
question of how precisely this mechanism operates to allow the agent to “stick to his
guns.”

Drawing upon Schelling’s work, Becker sums up the Side Bet strategy as follows:

The committed person has acted in such a way as to involve other interests
of his, originally extraneous to the action he is engaged in, directly in that
action. By his own actions prior to the final bargaining session he has
staked something of value to him, something originally unrelated to his
present line of action, on being consistent with his present line of
behavior. The consequences of inconsistency will be so expensive that
inconsistency in his bargaining stance is no longer a feasible alternative.22

Such, then, is the general mechanism of the Side Bet.

Some care must be taken about terminology. While Schelling is explaining the
ability to remain steadfast in one’s action through an irrevocable secondary
“commitment,” Becker is describing this secondary commitment as ensuring the
performance of some original “commitment,” whether it occurs in the bargaining process

(emphasis in original).

32-40 at, 35.
or elsewhere. Becker is seeking a general theory of the binding force of "commitment" to action, and he looking to Side Bets for an explanatory model. My focus is similar to that of Becker, except that I am not interested in the motivated behavior as an example of "commitment" in Becker's sense. Rather, I am trying to understand the mechanism of Output Willpower through an analogy to Side Bets as advocated by Ainslie. This naturally leads to three issues of concern: (i) what is staked in the relevant Side Bet, (ii) how is this staking accomplished, and (iii) what makes this staking irrevocable.

Significantly, Becker's model doesn't require the conscious staking of a Side Bet. Rather, the stake in the Side Bet can be "placed" merely as a result of circumstantial factors neither intended nor foreseen by the agent. He provides a kind of survey showing how membership in organizations and cultural groups, as well as one's culturally shaped expectations, give rise to intended and unintended "Side Bets" (if we are willing to construe the term so broadly). Becker gives the following example of a "Side Bet":

A person sometimes finds that he has made side bets constraining his present activity because the existence of generalized cultural expectations provides penalties for those who violate them. One such expectation operates in the area of work. People feel that a man ought not to change his job too often and that one who does is erratic and untrustworthy. Two months after taking a job a man is offered a job he regards as much superior but finds that he has, on the other side, bet his reputation for trustworthiness on not moving again for a period of a year and regretfully turns down the job. His decision about the new job is constrained by his having moved two months prior and his knowledge that, however attractive the new job, the penalty in the form of a reputation for being erratic and unstable will be severe if he takes it. The existence of generalized cultural expectations about the behavior of responsible adult males has combined with his recent move to stake his personal reputation, nominally extraneous to the decision about the new job, on that decision.²³

²³ Ibid. at 36 (emphasis in original).
As with the forgoing example, Becker generally offers examples in which the stakes in the Side Bet are simply the natural consequences (amongst which I include sanctions imposed on the basis of cultural determined rules) of the agent's primary decision. In other words, the agent does nothing to actively create an "irrevocable change in the payoff structure." Circumstances, rather than the agent, establish the terms of the Side Bet, and these circumstances can operate through whatever causal channel is appropriate in the particular case. Thus, Becker has no general theory of how a Side Bet is effected or what, aside from general causal necessity, makes it binding. Similarly, he does not have any general theory about the specific kinds of stakes that figure in Side Bets.\(^{24}\)

Ainslie and Haslam explain these Public Side Bets and draw an analogy to Personal Rules as follows:

This means of commitment [by Personal Rules] is an example of side betting, because it works by putting additional reward at stake in each individual choice. Schelling ... and Becker ... have described side betting as a way people sometimes commit themselves to a course of action by arranging to forfeit things of value—money, reputation, freedom—if they do not make a particular choice in the specified direction. For instance, a person may commit himself not to tell lies by cultivating a reputation as a truthful person, which would be lost if he were caught lying. In the examples of Schelling and Becker, however, the side bets are held by other people, who exact the penalty if the person does not behave as he said he would. Public side bets like those are a case of an extrapsychic commitment device. In the interdependent series of behaviors described earlier, no one need hold the bet; if he perceives himself not to have waited for the large reward, he will automatically pay the forfeit, and at once, by losing the expectation that he will wait for similar rewards in the

\(^{24}\) Schelling focuses on a less tenuous analogy to ordinary betting in explaining behavior in that he focuses on stakes one intentionally places on one's future behavior in order to insure that one will follow through on a given course of action. However, Schelling, like Becker, has no general theory about what kinds of stakes those are, how the staking is to be accomplished or what makes such staking irrevocable. Thomas C. Schelling, The Strategy of Conflict, (Cambridge: Harvard University, 1960, reprint 1980), 21-46.
future. The dieter who has, in effect, bet himself that he will not eat candy, pays for a lapse by a fall in his own expectations.25

The analogy is easily understood, but the basis for the irrevocability of the Side Bet requires closer analysis. However, before proceeding to that analysis, a few preliminary points must be made.

As an initial matter, some caution must be exercised with respect to Ainslie’s distinction between Public Side Bets and Private Side Bets. Becker, at least, is not committed to the idea that the stake in a Side Bet must be held by another agent, nor to the idea that enforcement must occur at the hand of another agent. Becker gives the following example:

Side bets constraining behavior also come into existence through the process of individual adjustment to social positions. A person may so alter his patterns of activity in the process of conforming to the requirements for one social position that he unfit himself for other positions he might have access to. In so doing, he has staked the ease of performance in the position on remaining where he is. ... [S]ome Chicago schoolteachers chose to remain in a lower-class school for the lengthy period necessary to reach the top of the list for a very desirable middle-class school. When the opportunity to make the move came, they found that they no longer desired to move because they had so adjusted their style of teaching to the problems of dealing with lower-class children that they could not contemplate the radical changes necessary to teach middle-class children. They had, for instance, learned to discipline children in ways objectionable to middle-class parents and become accustomed to teaching standards too low for a middle-class school. They had, in short, bet the ease of performance of their job on remaining where they were and in this sense were committed to stay.26

In this example, the stake in the Side Bet obviously is not held by another person, nor is the mechanism of enforcement through another person.


Further, the distinction between Public and Private Side Bets cannot be based upon whether the stake in the Side Bet is "internal" or "external." The Chicago schoolteachers in the example are staking "ease of performance," which is about as "internal" or "private" as any stake is likely to be. Ainslie's mechanism on the other hand, can stake any manner of reward in the Side Bet including external events, actions, or states of affairs. On a variation of the TV-watching example, the agent may stake a number of days of effectively completing various discrete objectives in his worklife ("external" rewards) against the pleasure of staying up late to watch TV. Ainslie's side stakes are not necessarily "internal," all that is internal is the utility or reward value of an act that is imputed to it because it purports to bear an evidentiary connection to the attainment of the relevant aims. This is just a general fact about evidentiary utility. For that matter, this imputation back of utility to an action operates in the same manner under a causal utility model.\textsuperscript{27} Setting aside Ainslie's claim to originality in regard to Private Side Bets as he characterizes them generally, we can look to the specific kind of Side Bet he proposes.\textsuperscript{28}

With these reservations established, I return to the general thrust of the Private Side Bet analysis offered by Ainslie and Haslam. Ainslie and Haslam note that the Side Bet need not be held by another person, since upon succumbing to temptation the agent would automatically revise his estimates of his ability to attain relevantly similar rewards in the future. The Side Bet is irrevocable because the agent's ability to claim or reclaim


\textsuperscript{28} For purposes of terminological simplicity, I will continue use Ainslie's term "Private Side Bets," at least for the time being, to distinguish it from the techniques operative in the kinds of cases that Schelling describes, and will not attempt to construct an ideal taxonomy of Side Bet mechanisms and techniques.
the stake is based upon his expectations, and his expectations are based upon his beliefs about his own future actions. These beliefs are based upon the conclusions the agent can credibly draw from the available facts. Under ordinary circumstances, his beliefs are not a function of his desires. Wishful thinking is possible, but this too is regarded a defect of rationality, and under ordinary circumstances is not a dominant factor in the formation of the agent's beliefs about his own actions. In order for the agent to believe something about his future actions, he must believe it to be true or at least credible. In trying to form beliefs that are true or credible, the agent must obey epistemic norms about how such beliefs are generated. These norms are outside of his control. Thus, the Side Bet is binding because it is based upon what the agent believes to be true or credible, rather than what he wants to be true. His finding a belief true or credible depends upon his believing it to be derived in a manner that is in basic accord with epistemic norms, which he finds as given and does not formulate according to his desires. In other words, epistemic norms are not formulated on the basis of the agent's desires, nor are the conclusions to which, under ordinary circumstances, they lead him. It is these facts about belief formation, and more specifically about the formation of the agent’s beliefs about his own future activities which make the Private Side Bet irrevocable.

Here we hit upon the core problem for Ainslie's idea of Prescriptive Personal Rules as a model of an intentionally made Side Bet. When the predictive nature of the Side Bet is absent because one prescribes one’s own Personal Rules, the irrevocability of the Side Bet is lost. When the Side Bet is no longer irrevocable, it loses its motivational force. To see this more clearly, suppose an agent were to make a Public Side Bet in the following way. He states that he would give up smoking for a month, or else pay $1,000
to the American Cancer Society. However, he does not entrust these funds to the care of the Cancer Society prior to his attempt to quit. He can certainly welsh on the bet, it is physically within his power to do so. In fact, all things being equal, he will welsh because if he ends up smoking he will have failed in his purpose to give up smoking, and it will be of no advantage to him \textit{at that time} to loose the additional reward of $1,000. Knowing this in advance, the Side Bet would be ineffectual in motivating his behavior.

The point here is still not as salient as it could be because our intuitions are apt to be that we might not welsh on the Side Bet. This is because the example is imperfect in that in imagining ourselves in this situation we are apt to implicitly import an additional stake in the Side Bet other than the $1,000. Our pride, self-respect or sense of our own honesty are likely to influence us in the direction of not welshing on the Side Bet. But note, these hidden or implicit stakes in the Side Bet are not revocable because they are based what we can credibly believe about ourselves in light of our actions and how they conform to the norms involved in those self-conceptions.\footnote{The topic of self-conceptions will be discussed further in Chapter Four.}

In order to isolate a purely arbitrary stake in a Side Bet, a stake which we put “on the table” purely by choice rather than by epistemic or other necessity, it is helpful to consider an example from Schelling that touches on the heart of the problem:

Both the Kidnapper who would like to release the prisoner, and the prisoner, may search desperately for a way to commit the latter against informing on his captor, without finding one. If the victim has committed an act whose disclosure could lead to blackmail, he may confess it; if not, he might commit one in the presence of his captor, to create the bond that will ensure his silence.\footnote{Schelling, \textit{The Strategy of Conflict}, at 43-44.}
The reason the example offered by Schelling is interesting is because it concerns a problem which surfaces in many areas of human endeavor, in situations far less dramatic. Situations often arise in which it is best for the agent now, and overall, to make it appear that he will take a certain action in the future. In the example, if the prisoner can convince his captor that he will not inform on him in the future this will be best for the prisoner now and overall. The best way to achieve this may be to make it true that the prisoner will not inform on his captor, or at least very likely that he will not, so as to make his statements to that effect credible. This is why the confession or perpetration of blackmail is contemplated.

To examine the issue more carefully, the reason making a credible commitment is an issue is because, ex ante, it is best for the prisoner, and satisfies his strongest current desires—i.e. his desire for survival, for the prisoner to make the kidnapper believe that he will not inform on him. However, _ex post_, after he is released, the prisoner will no longer have this incentive. Once out of danger, the prisoner has a strong incentive to inform on his former captor in order to get revenge, to see that justice is done, to prevent the kidnapper from kidnapping again, or for other such reasons. But the kidnapper will know this while the kidnapping is in progress, hence the need for, and difficulty in finding, a credible means for the prisoner to commit himself to not later inform on his former captor. More generally, the situation here illustrates the point that it may best satisfy the agent’s interests _now_ that he take a course of action in the future; but at the point in time at which it will be possible for him to take that course of action it may best satisfy his then-current interests to behave in a manner inconsistent with that same course of action. It is the recognition of this pattern of changing incentives and changing desires that gives
rise to the search for the means of making a credible commitment on the part of the kidnapper and his prisoner. For they know now that, all things being equal, the prisoner will later inform on his former captor in accord with his then-current strongest desires unless he finds some means of making a credible commitment now that he not do so. A good way to make such a primary commitment credible is to somehow make it real or binding on the prisoner so that it binds his future action.

The main point of this analysis for the purposes of the current argument is that, absent some method of making a binding commitment, the agent will in the future act according to what are at that time his strongest desires. He can make a "decision" now to act otherwise in the future, but all things being equal, he cannot now make a decision about what to do that will survive his future changes in motivation. His future action will be determined by the motivations he has then, not the motivations he had in the past. This is true even if his present strongest desire is for the endorsed reward and he recognizes that his future desires represent the specious reward.

But exactly the same reasoning applies to the secondary commitment forming the Side Bet. It might be in the agent's current interests now that the commitment he makes in a Side Bet be a commitment that he follows through on in the future. But this fact alone does not make the Side Bet binding. Indeed, the point of the Side Bet is to make the primary commitment binding, if the Side Bet itself is not binding, then it adds nothing to the agent's efforts to determine his future behavior. Consider again the kidnapping scenario. Suppose that instead of committing an act which would allow the kidnapper to blackmail him, the prisoner instead promises that if he (the prisoner) later informs on the kidnapper, then he (the prisoner) will, at that time commit some act that will lead to his
(the prisoner’s) being sent to jail. This would hardly be convincing to the kidnapper, because it would not be in the prisoner’s interest at the time at which he (the prisoner) might inform on his captor to do some additional harm to himself if he did so inform the authorities. Indeed, at that time it would seem a truly bizarre thing to do.

In sum, the Side Bet is not motivationally effective unless there is some additional factor that makes it binding. If the person making the Side Bet is free to welsh on the bet in the future if he turns out to lose it, and there is no collateral consequence for him doing so, then the Side Bet will hardly be binding. Knowing he can simply “welsh” when the time comes to “pay up,” winning\textsuperscript{31} or losing the “bet,” if indeed such a revocable arrangement can be called a “bet,” will not form a stake in the agent’s behavior at the time at which he carries out the performance on which the bet is staked. This is to the agent’s detriment if he is making a Side Bet to prevent himself from acting on a specious reward, since the Side Bet will offer him no means to overcome temptation. This follows whether the Side Bet is public or private. A Side Bet is not binding merely because it is in the agent’s interest at the time he makes it for it to be so. I will call the implicit assumption that the Side Bet is binding merely because it is in her interest for it to be so the “\textit{Bootstrapping Problem}.”

With this analysis in mind, consider Ainslie’s sole direct argument for

Prescriptive Personal Rules. Again, he uses the analogy to a Side Bet.

\ldots [T]he whole series of larger rewards [is] to be forfeited if, at any choice point, the specious reward were not avoided. It would not be necessary

\textsuperscript{31} It must be recalled that “winning” in some such cases just means “not forfeiting the stake in the Side Bet.”
for anyone to hold the bet, since the mere knowledge that this bet was necessary to avoid the specious rewards would make it binding.\textsuperscript{32}

Hence, Ainslie believes that for this reason one can construct Personal Rules that are motivated by an “arbitrarily propose[d]” Side Bet.\textsuperscript{33}

This attempt to bootstrap oneself into a successful Side Bet simply will not work. If no one, or no thing, is “holding the bet” or, more precisely, “enforcing the bet,” then the bet will fail. As Ainslie’s analysis of Predictive Personal Rules shows, it may not be a person who “enforces the bet”—i.e. who makes it binding. A bet may be made irrevocable because, for instance, epistemic norms which guide belief may make the bet irrevocable because one’s action is taken to be a reliable predictor of future action. However, “mere knowledge that the bet was necessary to avoid the specious reward,” as I have argued at length, simply is not sufficient and would not make it binding.

Consider again the Personal Rule that I will shine my shoes every morning before breakfast. Ainslie describes this as follows:

[If a person wanted to commit himself to keeping his shoes shined, he might make a rule that he would shine them every morning before breakfast. He would be betting whatever the expectation of having shiny shoes was worth to him against his tendency to skip shining them on any particular morning, since if he skipped one morning he would be just as apt to skip the next.]\textsuperscript{34}

But why would the agent’s expectation of having shiny shoes be staked on his shining them on a given morning? By hypothesis, his shining his shoes on this morning

\textsuperscript{32} Ainslie, “Specious Reward,” at 479 (emphasis in original).

\textsuperscript{33} Ibid. at 480. I rely upon Ainslie (1975) for an explanation of Prescriptive Personal Rules since Ainslie, in his later and more developed work, simply refers the reader to this article when the need arises to explain how such Side Bets can be “arbitrarily propose[d].” See e.g. Ainslie & Haslam “Self-Control,” at 188; Ainslie Picoeconomics, at 152.

\textsuperscript{34} Ibid. at 481.
is no predictor as to whether he will shine them in the future before he makes his rule. So the rule must do the trick. But what makes the rule binding? As I have argued at length, it cannot be the mere fact that it would be advantageous if such a bet were binding.

Suppose the agent fails to shine his shoes on a given morning. He has lost the bet. This should mean that, barring some special circumstances where the terms of the bet are reformulated,\textsuperscript{35} he should forfeit the stakes. That is what give the bet its motivational influence. In this case, the agent suffers a loss in his expectations that his shoes will be shiny in the future. But why should he suffer such a loss in expectations if shining his shoes on this day was no predictor of whether he would shine them in the future—at least not before the rule came into play? There are only two choices. Either creating the rule automatically makes this day’s performance a predictor of future performances, or the agent independently enforces the bet himself. In either case, if such results were sure to follow, the agent would expect to pay the price in terms of lost expectations at the time he failed to shine his shoes. Hence, the bet would be motivationally efficacious at the time of decision as to whether he should shine them.

Consider first the case where the agent enforces the bet independently, so to speak. This would mean intentionally not shining his shoes on some substantial number of future mornings. But why would he do that? Ex ante, this would be a silly thing to do. This would be like the kidnapper’s prisoner doing some act for which the kidnapper could blackmail him after the kidnapper has released him and after he told the police who the kidnapper was. Ex ante, the agent would be aware of this fact, and so (absent some

\textsuperscript{35} Ibid. at 481-82. It cannot be that the general rule that Side Bets would simply be reformulated to cover only future cases subsequent to this failure in order to preserve the preserve the power of the relevant Side Bet for future use. For if every failure could be so excused, the Side Bet would lack force. Ainslie acknowledges this. Ainslie “Specious Reward,” at 481-82.
additional motivational explanation) the Personal Rule will never be credible and would, therefore, lack motivational force.

At this point, Ainslie himself asks why a person would abide by the terms of a self-enforced Side Bet in the form of setting aside some reward which one will grant oneself only if he should complete some target task:

Self-reward is an intuitively pleasing strategy until one asks how the self-rewarding behavior is itself controlled. A subject does not actually recruit additional reward by planning to delay a cigarette until he has finished a difficult task. On the contrary, he sets himself a second task: He must both defer smoking and work on his original task on the basis of the same differential reward that has always confronted him.36

The discussion following these remarks is not very clear. However, Ainslie appears to be trying to explain the functioning of self-reward in terms of binding Private Side Bets, rather than explaining Private Side Bets in terms of the agent’s ability to voluntarily enforce his own Private Side Bets through self-reward or self-punishment.

It is just as well that Ainslie does not take the latter route, for if he did there would be two very serious consequences for his theory. First, Prescriptive Private Side Bets would lose much of their interest since they could simply be assimilated to the model of self-reward. They would just be one subset of cases of self-reward not particularly more interesting than any other subset. Secondly, this move would further undermine the effort to make Personal Rules an explanation for Output Willpower. For if an agent voluntarily engaged in self-reward or self-punishment in order to enforce his Side Bets and make them credible, such a technique as an explanation of Output Willpower would run seriously afoul of the Device Objection. In general, there is no advance here in solving the Linkage Problem.

36 Ibid. at 485 (cites omitted) (emphasis in original).
Having dispensed with the idea of voluntarily enforced Prescriptive Private Side Bets, the remaining option is that formulating the Personal Rule somehow makes it self-enforcing. The only way Ainslie has shown us around the Bootstrapping Problem thus far is by showing that there may be a predictive connection between one choice point and another. Perhaps Ainslie holds that Prescriptive Personal Rules can establish this predictive connection "arbitrarily," and once established, it is as binding as any Predictive Personal Rule. The question remains: How? After all, I cannot "arbitrarily" establish a predictive relationship between any two events. I cannot, for instance, simply decide that if it rains tomorrow that will be valid evidence that I will be elected president of the United States.\textsuperscript{37} It must be then, not that the agent decides that the two decision points will be linked by a predictive relation by fiat, so to speak, but, rather, something about the agent's decision to act, or his conception of it, other than some fantasized predictive connection or the mere fact that it would be advantageous to him for the decisions to stand or fall together, that binds the decisions points together. However, Ainslie has not explained what this factor could be, and so we must move beyond his theory of Personal Rules in seeking an explanation.\textsuperscript{38}

\textsuperscript{37} At one point, Ainslie argues that an agent will not welsh on his Side Bet by changing the terms such that, post violation, behavior which violates the rule gets redefined so that the apparent violation is no longer deemed a violation. Ainslie suggests that welshing on a Private Side Bet in this manner may challenge the credibility of other, similar Side Bets, and that the agent will not want to pay the price for such loss in credibility in terms of diminished future expectations. Ainslie "Specious Rewards," at 482. However, it is not yet clear how any Prescriptive Personal Rule could have credibility to begin with. Therefore, with regard to Prescriptive Personal Rules, this argument produces a vicious regress. For if there is nothing connected with this Side Bet that makes it binding other than the fact that welshing would undermine future Side Bets, the presupposition is that those future Side Bets are, for some reason or other, credible and binding. But what makes these other Side Bets credible and binding? It can only be their potential impact on still other Side Bets, and so on ad infinitum. There is no indication as to what grounds this great web of Side Bets in the first place.

\textsuperscript{38} Ainslie offers a further analysis in terms of prisoner's dilemma that may be, in part, another attempt to defend the idea of arbitrarily constructed Side Bets. However, the analysis is fundamentally incoherent and fails to offer any support for Prescriptive Personal Rules. See Appendix A.
IV. Further Variations on a Theme
   A. Introduction
      In this section I will look as several further variations of the Personal Rules
      Mechanism not envisioned by Ainslie himself. I have stressed that one of the key
      problems plaguing Personal Rules is the Linkage Problem—i.e. the connection between
      the decision to act and the engagement of willpower. Nozick and Bratman have sought to
      solve this problem by elaborating ways in which a sequences of decisions can be made to
      stand or fall together, so as to bring the power to connect these decisions within the
      agent’s control. I will argue that the strategies they employ to achieve this effect do
      nothing to solve the Linkage Problem. Moreover, their accounts fail as an explanation of
      willpower on a variety of other general grounds. Finally, I propose a distinct new
      mechanism called “Will-Based Side Bets” and argue that it shows something about the
      kind of connection willpower should have with the decision to act, but doesn’t get it quite
      right. Additionally, Will-Based Side Bets suffer from a number of serious problems
      including an inability to reflect some rough proportionality between the agent’s resolve to
      act willfully and a concomitant boost in motivation.

   B. Nozick’s Version of Personal Rules
      Without providing any textual support from the work of Ainslie, Nozick explains
      the workings of Personal Rules along lines quite different from those we have
      investigated thus far. Nozick’s reading of the relevant features of Ainslie’s Personal
      Rules is captured in the following paragraphs:

      The mark of a principle (“Never eat snacks between meals”: “Never
      smoke another cigarette”) is that it ties the decision whether to do an
      immediate particular act (eating this snack, smoking this cigarette) to the
      whole class of actions of which the principle makes it a part. This act now
      stands for the whole class. By adopting the principle, it is as if you have
      made the following true: if you do this one particular action in the class,
      you will do them all. Now the stakes are higher. Tying the utility of this
act of snacking to the disutility of all those acts of snacking in the future may help you to get through the period ... of temptation; the utility for you now of this particular snack is altered. This snack comes to stand for all snacks, and at this early point the current utility of being thin or healthy later far outweighs the current utility of those distant pleasures of eating; the current disutility of poor health or a poor figure becomes a feature of the currently contemplated particular act of snacking. ...

By adopting a principle, we make one action stand for many others and thereby we change the utility or disutility of this particular action. This alteration of utilities is the result of exercising our power and ability to make one action stand for or symbolize others. Violating the principle this one time does not necessitate that we will always violate it: having this snack does not necessitate that we will become continual snackers. Before we adopted the principle it was not true that doing the act this one time would involve doing it always. Adopting the principle forges that connection, so that the penalty for violating the principle this time becomes the disutility of violating it always...

What makes Nozick’s account interesting is that, whereas Ainslie speaks of a given instance of temptation as a kind of “test case” providing evidence about the likelihood of similar future failures of will, Nozick speaks of that action symbolizing such future failures of will. I will call this “Symbolic Personal Rules.”

Nozick then goes on to consider whether violating such a principle thus established merely “... affect[s] our estimate of how likely ... repetition [of the same type of violation] is. . .” or whether such a violation also does “... actually affect the future ... [i.e.] makes it more likely that we will repeat the action.” Nozick concludes that violating a principle does make it more likely that we will succumb to temptation through a variety of causal mechanisms including those previously discussed in this dissertation.

---


40 Ainslie, Picoeconomics at 152, note 4.

41 Nozick, The Nature of Rationality, at 19 (emphasis in original).
Thus, on Nozick’s version of Personal Rules, they tend to be self-fulfilling, just at they are on Ainslie’s view.

Returning to the Linkage Problem, Nozick’s account lacks any explanation of just how this symbolic connection is initially established. The first question is whether an agent can simply choose to activate the symbolic connection between one action and other actions or events. This would make the connection a result of a distinct choice in much the same manner as Ainslie theorizes is the case Prescriptive Personal Rules. The text is somewhat ambiguous on this point. In some instances, Nozick appears to hold that the symbolic link cannot be forged voluntarily,\(^{42}\) while in other places he seems to suggest that it can be.\(^ {43}\) Unfortunately, Nozick offers no examples of the latter sort of case, except that he appears to presuppose that it can be done in his description of the functioning of Personal Rules. I think this is an assumption open to serious doubt.

Nonetheless, suppose such a symbolic link could be intentionally forged by an agent and that Symbolic Personal Rules are constructed in the manner Nozick suggests. The first point that this assumption draws out is that, under Nozick’s account, what is important for agents is what we might call the “rule in place,” as opposed to the act of rule making. Consider Nozick’s example of a rule: Never smoke another cigarette. What provides the added motivation is the connection between one instance of smoking and potential future instances of smoking via the rule. From the standpoint of the added motivation, the mere fact that I made the rule one year ago, rather than five minutes ago,

\(^{42}\) See e.g. Ibid. at 28 (discussing the implications of the fact that a given “... symbolizing connection always holds ...”).

\(^{43}\) See e.g. Ibid. at 30 (“A large part of the richness of our lives consists in symbolic meanings and their expression, the symbolic meanings our culture attributes to things or the ones we ourselves bestow.”) (emphasis added); Ibid. at 32 (“We live in a rich symbolic world, partly cultural and partly of our own individual creation... .”)
does not necessarily lower its motivational impact. Smoking one cigarette now implicates any number of future episodes of cigarette smoking once the symbolic link is forged via a principle.44 There is no special motivation attached to the act of formulating the rule for the first time under this account. It may be that one is particularly excited about having formulated a rule—psyched up, so to speak, about one’s new program of action. But such would be an adventitious phenomenon, unrelated to the operation of the rules mechanism itself. The same, of course, goes for reminding oneself of the rule at a later date. Having attention focused on the rule may yield a motivational affect, but this is a separate process, describable under the general mechanism of control of attention or providing oneself with certain stimulating signals to activate the right motivational “nodes” as some theorists would have it.45 In sum, the rule in place is the name of the game. This will become clearer in what follows.

These considerations draw out a point that was hidden by the ambiguity of Ainslie’s account of Personal Rules due to his shuttling back and forth between rules and Side Bets metaphors. Classically, establishing a Side Bet happens once, as with Schelling’s buyer, and this bet motivates the action until the bet is either won or lost within the bounds of a single episode, so speak. A new bet must be placed to cope with the next situation where a Side Bet is going to be worthwhile. Predictive Personal Rules consist of what amounts to a series of analytically distinct Side Bets. At each decision

44 Ibid. at 19-20. Of course, circumstances external to the rule mechanism may change over time such that its efficacy is diminished, but this is a separate issue. For instance, one year after I formulate my rule my doctor may inform me that I will die within three months. At this point, my no-smoking rule is not going to seem all that important to me. (In fact, Nozick stresses that as one invests more in obeying a rule over time, the consequences of breaking it may be yet more grave. “For how likely is it that you will continue to adhere to another principle if you couldn’t manage to stick to this one despite so much effort?”

45 See Chapter Two.
point some collection of future decisions are staked on the present choice. This
collection, as it figures in the instant bet, will necessarily be different from that figuring
in the previous bet. This is because one of the “future decisions” figuring as a side stake
in the last Side Bet now figures as the primary decision at issue in the instant Side Bet
situation. One has moved one step further down the line of decision points in the relevant
series. Even though two bets in the series may look alike, they are not the same. A new
“bet” is made with each decision. By contrast, a rule is made once and followed or not
followed on subsequent occasions.

I think Ainslie has gone wrong in trying to assimilate the model of a structured
sequences of Side Bets to the model of a rule, and I think Nozick has followed him in this
error. I suspect this is why one tends to look in vain for convincing examples of the
“rule” in Ainslie’s examples of Personal Rules. Nozick tries to make this analysis work
a little better, but his account is sketchy and dependent upon a very questionable idea of
symbolic connection, such that it is far from clear that it represents any advance.

One serious problem is that Nozick’s proposed symbolic connection between an
actual decision and action now and anticipated actual decisions and actions in the future
is that it can’t seem to do the job that Nozick sets for it. Nozick makes plain that the
symbolic connection—of itself—has no causal or predictive bearing upon future actions.
Hence, if the connection is only symbolic, those future decisions do not actually stand or
fall with this one, aside from adventitious secondary effects, even in the agent’s own
understanding of the situation unless we assume that the agent is grossly confused about
the nature of the connection. And it is hard to square his being so confused about the

123
nature of the symbolic connection when it is supposedly one that he himself intentionally forged.

If the decisions in the iterated series do not actually stand or fall together, it becomes difficult to see why one needs a sequence of genuinely anticipated future decision points, as opposed to merely imaginary or hypothetical decision points, in order for Nozick’s mechanism to work. The real import of this mechanism is that the agent may say to himself, “If I violate this principle here, then this is the same as (i.e. “stands for”) violating the principle in these future cases as well.” In other words, violating the principle here stands for the fact that I would violate this principle generally, or in any number of other cases. If you like, it shows that I am the sort who violates this principle. Genuinely anticipated future cases where I will have to decide whether to abide by the principle or not really serve as little more than convenient illustrative cases on this analysis. For this reason, the principle at work never really changes. The principle is always the same over a series of choices because it needs only imagined future cases to illustrate its point. By contrast, a new Side Bet is made in every case.

In sum, I think there is a fundamental tension in Nozick’s account between the idea of actual future decisions being at stake and a merely symbolic connection between them and the present choice. This tension can only be resolved by completing the move toward the symbolic and supposing the connection to be with hypothetical, as opposed to actual, future decisions, though genuinely possible future choices may, so to speak, serve as particularly vivid examples for purposes of the imagined future decisions. Once this move is made, I think that we might have something that looks like a more faithful account of principles considered as one of the mechanisms that motivates the agent, and a
better grasp of why it is that he would abide by a principle once it is established.

However, at this point it is plainly evident that Ainslie’s mechanism has been left far
behind and with it the idea of constructing an account of Symbolic Personal Rules.

Given this reconstructed version of Nozick’s account, does this new apparatus,
which I will call a “Principle,” leave us in any better position to resolve the Linkage
Problem? I do not think so. Awareness that an action violates a Principle that an agent
holds simply reveals to him the broader consequences of violating that Principle. It
shows that one would be willing to commit violations (plural) of that Principle. One
might take this a bit farther and say, it shows that one is the sort who would do these
kinds of things. In abiding by a Principle, one is merely acting in light of the
consequences in this broad sense. The potential consequences of the agent’s decision
about which action to take -- i.e. in accord with the Principle or not in accord with the
Principle -- exist independently of that choice and provide the motivation that will
ultimately determine that choice. At the point of the decision to take action, the agent
merely responds reactively to the payoff structure already in place due to his prior
adoption of the Principle. His motivation does not arise from his resolve to undertake
action.

The agent is likely to cognize this in terms of wanting to avoid certain ethical or
prudential evaluations of his behavior, of being a “bad,” or conceivably a “weak” person.
He passively responds to these incentives in place. Although the agent may not be able
to fully articulate the point, there is a clear sense in which he will be propelled by these
considerations, by circumstances outside his control, and he will be aware of this. This is
to be contrasted with James’ paradigm case of the employment of Output Willpower.
There, the agent had tallied up all the pros and cons in terms of consequences for deciding either way, but was still unmoved until such time as he assented or resolved to act, thereby bringing the added motivation of willpower into play.

It is important to note that this objection is purely general. Even if we drop the idea of a Principle altogether, the symbolic connection in and of itself gets us no further.\footnote{Nozick never makes this move, presumably because he takes himself to be presenting a better version of Ainslie's mechanism and finds the idea of the connection between willpower and a Principle particularly apt, perhaps due to vaguely Kantian considerations.} Consider Nozick's discussion of the symbolic connection more generally.

One mark that it is an action's symbolic connection to an outcome that plays a central role in the decision to do it, rather than the apparently causal connection—I am thinking of cases where the agent does not think the action itself intrinsically desirable or valuable—is the persistence of the action in the face of strong evidence that it does not actually have the presumed causal consequence. Sometimes a person will refuse to look at or countenance this evidence or other evidence about harmful consequences of the action or policy. (On these grounds, one might claim that certain antidrug enforcement measures \textit{symbolize} reducing the amount of drug use and that minimum wage laws \textit{symbolize} helping the poor.) A reformer who wishes to avoid such harmful consequences may find it necessary to propose another policy (without such consequences) that equally effectively symbolizes acting toward or reaching the goal. Simply halting current action would deprive people of its symbolic utility, something they are unwilling to let happen.\footnote{Ibid. at 27 (emphasis in original).}

Precisely the same considerations apply here as apply to Principles. This is choice in light of consequences, not motivation arising through resolution to act.

Suppose, in fact, that we knew what the end or reward was that provided the motivation at work in Output Willpower (for under BMA such an end must exist). The problem remains that if one's action symbolized that thing (or event or process or state of affairs), then this fact would figure as a consequence at stake in one's decision antecedently, that is, regardless of one's decision to act or one's resolution to undertake
action, at least unless some further special linking connection could be established between the decision to act and the creation of the symbolic connection. The symbolic connection alone does not get the job done. Nozick does not provide any suggestion as to a link between the decision to act and the concomitant creation of a symbolic connection, nor am I aware of any other theorists having done so. Thus, Nozick’s account of symbolic utility does not solve the Linkage Problem because the theory of symbolic utility, standing alone, does not account for the connection between the decision to act and the engagement of Output Willpower.

Supposing that the symbolic connection could be established voluntarily and on a case-by-case basis is of no help for familiar reasons. With respect to the possibility of intentionally forming a symbolic connection, the most plausible scenario is to suppose that one could decide whether or not one would regard a choice as falling under a Principle, or more generally that one could decide that a given action or its consequence symbolizes some further action(s) or consequence(s) respectively. Here, one could elect how he chooses to frame the situation of the decision.48 The employment of such a framing technique via a separate decision is, however, subject to the Device Objection. Once again, the agent would be making a distinct decision to cause himself to have the motivation to achieve some end. The motivation does not arise as a natural concomitant of his resolution to take action.

Hence, as yet, we have no reason for thinking that a symbolic connection in and of itself is of any particular value in solving the Linkage Problem. As a hypothesis it may

offer some valuable insight into a distinct motivational mechanism, or set of mechanisms, that allows for a little more of the flexibility agents seem to have in constructing their own rules of conduct than did Ainslie’s Personal Rules. However, the hypothesis that there may be a symbolic connection between actions, or between actions and their consequences (real or hypothetical), does nothing to solve the Linkage Problem. 49

C. Bratman’s Version of Personal Rules

Bratman also attempts a reconstruction of Ainslie in an effort to find a way in which a series of decisions can, somehow, stand or fall together. 50 In so doing, Bratman relies upon his theory about the nature of intentions. As interesting as this move is, it

49 It should be noted that Principles do not provide an answer to the Ordinary motive objection or the Multiple Motives Objection. With respect to the latter objection, Principles rely upon the idea of a structured series of decision points. Specific Principles encompass specific classes of ordinary behavior associated with the attainment of ordinary rewards. Different Principles subsume different classes of decision points and these different classes will implicate different kinds of rewards at stake in the relevant choices. Hence, multiple Principles imply multiple motivations for action. With respect to the former objection, the rewards at stake involve either some symbolic realization of the ordinary rewards contingent upon the choices in the structured series, or the inherent reward or “punishment” involved in obeying or violating a specific Principle. Hence, the motivations are rather ordinary, and apparently not the special motivation involved in Output Willpower. A reformulated One-Shot Objection could also be constructed to the effect that when, for instance, the swimmer is preparing to dive into the cold water, it seems most plausible that he neither imagines future divings-in nor does he contemplate the “Principle” that the act of diving in stands for. However, it clear that the Unraveling Problem will not have any bite on Principles given that the imagined series of decision points can presumably be as long as the agent chooses to visualize it as being.

50 One of Bratman's chief grievances with Ainslie's Private Side Bet model is that it seems to confuse acts that cause consequences with acts that evidence consequences. More specifically, it looks as if the agent is doing an act that ordinarily serves as evidence of an outcome in order to make it more likely that the outcome will occur—but the ability to make the outcome occur seems to presuppose a causal connection. Thus, the agent goes to bed early in order to provide evidence that he will go to bed early on future occasions, but it is hard to see how this could be evidence of such unless there is an underlying causal connection. Bratman frames his criticism in terms of the famous philosophical puzzle: Newcombe’s Problem, Michael E. Bratman, “Planning and Temptation,” Faces of Intention: Selected Essays on Intention and Agency, (Cambridge: Cambridge University Press, 1999a): 35-57. Ainslie has long been aware of this objection and its supposed connection to Newcombe’s Problem, Ainslie, Picoeconomics, at 203-205. See also Ainslie “Studying Self-Regulation,” at 16-20 (responding to Bratman). However, since I have found Predictive Personal Rules wanting as an explanation of willpower on other grounds, I will not take up this issue. It is interesting that Nozick does not take up this issue in his discussion of Ainslie’s Personal Rules despite the fact that in the chapter after his discussion of Personal Rules Nozick launches into a discussion of Newcombe’s Problem. Nozick The Nature of Rationality.
takes us no closer to solving the Linkage Problem specifically or explaining Output Willpower in general.

Bratman argues that when faced with a series of choices of the sort Ainslie describes, we should not assume that one can only choose between the two rewards at stake in the choice immediately at hand. Rather, "[w]hy not take more seriously the idea . . . that I really do choose between . . . sequences or policies . . ."51 In other words, one chooses between for instance, (i) a policy or plan of staying awake later and watching TV on the relevant series of future nights and (ii) a plan or policy of going to bed early on the relevant series of future nights.

Bratman’s explanation of how one does this is based upon his general theory of intentions. On Bratman’s view, “[i]ntentions are, so to speak, the building blocks of . . . plans; and plans are intentions writ large.”52 Thus, intentions serve both to coordinate an agent’s activities over time and to coordinate his interaction with others in carrying out joint or related activities.53 In many cases, such as the cases of temptation we have been describing, it will be advantageous to the agent if he can formulate such an intention or plan ex ante and stick to it in the face of temptation. Bratman describes what he calls the “volitional dimension of commitment” in explaining how we stick to such plans as follows:

... Both intentions and desires are, but ordinary beliefs are not, pro-attitudes. Pro-attitudes in this very general sense play a motivational role: in concert with our beliefs they can move us to act. Thus, both the desire


52 Ibid. at 8.

53 Ibid. at Chapter 3; Bratman Faces of Intention, Part Two.
to go to the Tanner Library and intention to go could, in concert with a belief that a certain bus will take me there, lead to my taking that bus.

But we may go on to distinguish between two kinds of pro-attitudes. To change examples, 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 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 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.

This example illustrates an important difference in the normal functioning of desires and intentions concerning what to do in present circumstances. Intentions are, whereas desires are not, conduct-controlling pro-attitudes. Ordinary desires, in contrast, are merely potential influencers of action. The volitional dimension of commitments involved in future-directed intention derives from the fact that intentions are conduct controllers. If my future-directed intention manages to survive until the time of action, and I have seen that the time has arrived and nothing interferes, it will control my action then. As conduct-controlling pro-attitudes my intention involves a special commitment to action that ordinary desires do not.\(^5^4\)

Bratman states further:

... Future-directed intention involves volitional commitment because of its relation to present-directed intention: if my future-directed intention manages to survive until the time of action, and I see that this time has arrived and nothing interferes, it will control my action then. Present-directed intentions have a special relation to action, and future-directed intentions are the sort of state that will have this special relation if they survive until the time of action is seen to have arrived.\(^5^5\)

Thus Bratman explains that intentions carry us through periods where we might otherwise behave differently because we don’t reconsider our plans—more specifically,

\(^5^4\) Ibid. at 15-16 (emphasis in original).

\(^5^5\) Ibid. at 108.
we don’t reconsider whether our aims are appropriate ones. Bratman provides a detailed analysis of when and why it would be rational to stick to such a plan and avoid reconsideration of it,\textsuperscript{56} and I do not take issue with these conclusions. The real question is: How do we do this?

This is crucial. After all, the mere fact that it would be rational for an agent to form and maintain such an intention in no way guarantees that he will do so, for that, as we have seen, that is the crux of the Bootstrapping Problem. Moreover, at the time Output Willpower is engaged, all rational considerations may have been brought to bear on the situation already without yielding the appropriate behavior on the part of the agent. In fact, \textit{ex hypothesi}, at the time of choice, all things being equal, the agent is most strongly motivated to make the current choice that would constitute making an exception Bratman’s rationally conceived intention or plan. So, again: what makes us stick to our plan?

Unfortunately, Bratman does not provide any specifics on this issue. He does suggests generally:

An instrumentally rational planner will have mechanisms and strategies of reconsideration that sometimes block reconsideration of prior intention in the face of merely temporary preference change … This means that prior intentions and plans can have an independent role in rational motivation: once in place they can sometimes rationally control conduct even in the face of temporary preference change to the contrary.\textsuperscript{57}

What Bratman is suggesting is that secondary mechanisms or techniques must come into play to preserve the intention. Under BMA, such techniques and mechanisms must not only block "reconsideration of prior intention," they must also prevent

\textsuperscript{56} See generally Bratman \textit{Intention}.

\textsuperscript{57} Ibid. at 53.
competing motives from being on balance stronger than the motivations in favor of the plan. Otherwise, under BMA we will act in contravention of the plan whatever our rational conclusions about it. One might imagine control of attention, for instance, having a role here in helping to insulate the plan from competing motivations or enhancing its motivational impact through attentional focus upon the plan or its consequences. While this structure might in itself be an idea well worth pursuing, it makes it plain that the concept of an intention does no motivational work in the operation of Output Willpower. Rather, it is some other unspecified mechanism or technique that provides the necessary motivation to “get over the hump” and stick to the plan. Moreover, whatever may be going on here, it sounds like Input Willpower rather than Output Willpower. The mechanism may, in addition, be subject to the Device Objection. This being the case, we will have to look for our explanation of Output Willpower elsewhere.

D. A Final Variation: Will-Based Side Bets

There is yet another move that can be made in trying to rehabilitate Ainslie’s Private Side Bet model in the face of the Linkage Problem. The idea is simple. I have gone to some lengths to describe the folk conception of Output Willpower. We can supplement Private Side Bets by making use of the folk concept in describing an agent’s

decision to act. Thus, when an agent *decides* to perform some action, whether he succeeds or fails in following through may appear to him to be a test of willpower. If he has the willpower to succeed in this case, then this is some evidence that he will succeed in relevantly similar future instances. Therefore, he will be motivated to prove that he can succeed in this instance in order to also prove that he has sufficient willpower to succeed in those future decisions. The predictive quality of his actions will tend to be self-fulfilling for all the reasons already given by Ainslie. In sum, the agent has made a binding Private Side Bet that springs naturally from his decision to carry out an action. I will call this a "Will-Based Side Bet."

The Will-Based Side Bet navigates between the opposite dangers of being either too voluntary or not voluntary enough. That is, on the one hand it does not go so far as to present a Side Bet that is set in motion by a deliberate choice that is wholly distinct from the decision to take the action. That would leave the account open to the Device Objection. On the other hand the Side Bet is not engaged willy-nilly, in advance of any decision by the agent so that it is merely an artifact of the structure of the decision points that lie before the agent. This seems an advance in solving the Linkage Problem.

While this may be a step forward in addressing the Linkage Problem, I have thus far glossed over a number of serious problems. The most important problems, for my purposes, are that the structure that this account proposes neither explains when willpower will be employed nor to what degree it will be employed. In the first two chapters I described Output Willpower as not being operative in every decision to take action that an agent makes. I also described the added motivation of Output Willpower as capable of varying degrees of application. In other words, Output Willpower is not
simply governed by an on/off switch that sets in motion a fixed quantity of motivation. Under the folk conception, the agent may use more or less willpower at a given time in addressing a given task. The structure of Will-Based Side Bets gives no indication as to whether or how much willpower will be activated by a decision. Since there is no determinate relation between a decision and the engagement of the Side Bet, success or failure bears no fixed evidentiary relation as to one’s “strength of will” or capacity to exercise willpower.

It is tempting to say that this relation is fixed by how much one “tries” to carry out the action. However, such an explanation cannot be cashed out in terms of magical volitions—trying explained as some peculiar mental “ghostly thrust.” Rather, “trying” must be explained in terms of the motivations actually employed that can operate to produce action. But this would seem to put the cart before the horse. The suggestion is that one monitors one’s operative motivations in order to see how much willpower one has engaged, checking up on oneself after the process is in motion. But on the folk theory, one gives the “heave of the will” as a part of intentionally undertaking the action. One knowingly and intentionally engages in willful action, and regulates the degree of motivation voluntarily. One does not discover *ex post facto* that he has engaged his will and then attempt an estimate of the degree of its employment.

Moreover, there is a fundamental confusion involved in the *ex post facto* account, even on its own terms. For it looks to the motivations already engaged to estimate how much one is trying, i.e. how much willpower is in play; but then it is this estimate which governs the strength of the predictive link between success in the present decision and success in relevantly similar decisions in the future. But this predictive connection is
what is supposed to give rise to willpower. That is, the evidentiary link is supposed to
explain willpower, not result from it. Thus, the theory is self-contradictory in suggesting
that the agent look to one set of motives to determine how much willpower he is exerting,
but explains his experience of exercising force of will through a second motivation or set
of motivations occurring later in the process.

Once could try to get around this problem by proposing that there is a feedback
loop in operation that makes the various motives indistinct. That is, one looks to one’s
efforts to see how hard one is trying, which implicates a certain amount of will at work,
which in turn increases motivation through the Side Bet Process, which visibly increases
the extent to which one is trying hard to take the action, which calls for an upward
adjustment in the stakes in the Side Bet, which visibly increases the extent to which one
is trying hard . . ., etc. In this way, the distinction between ordinary motives and the
boost in motivation associated with willpower could become blurred, since the whole
“cycling up” process could occur very rapidly. The agent might well experience an
otherwise unaccounted for boost in motivation that he can only attribute to willpower.

However, even assuming that this feedback process could work, it faces several
further objections as an explanation of Output Willpower. First, there is no clear
indication as to how this cycling up of motivation is to be controlled. It is difficult to see
how it could be achieved without opening the door to the Device Objection. Second,
without having anything else to go on, this automatic cycling up of motivation sounds
like a recipe for compulsive behavior rather than willpower. There seems to be nothing
to stop this feedback process from generating motivations well out of proportion to the
agent’s estimation of the importance of the task at hand. Even if the mechanism does not
"overshoot" in terms of the amount of motivation appropriate, it may well "undershoot."
The connection between the agent's estimated importance of the task and his
commitment to carrying it out, on the one hand, and the motivation level produced by this
mechanism, on the other, is wholly incidental. I will call this problem of accounting for
some proportionality between the level of one's resolve to undertake action by force of
will and the level of the added motivation produced in response the "Proportionality
Problem."

Finally, the link between willpower and the decision to act remains incidental on
the ex post facto view. Whatever set in motion the first-order motivations to which one
looks in order to estimate his willful efforts, be it rewards or aims of any kind, these
motivations were not set in motion as a result of the agent’s decision to act, rather they
motivated that decision in the first place. As for the motivations generated by the
hypothesized evidentiary connection that results from the ex post facto estimate of
"willpower" engaged, neither are these connected in any way with the decision to take
action. Rather, it is simply another automatism, or at best a device guided through
control of attention. In the end, the structure outlined thus far grinds to a halt on the road
to solving the Linkage Problem.

V. Conclusion

In this chapter I have argued that Ainslie's account of Personal Rules has failed as
an explanation of willpower. Predictive Personal Rules, when not employed
intentionally, are not the results of any choice on the part of the agent, and so cannot
explain for willpower. Even when intentionally employed, Predictive Personal Rules fail
to explain willpower for several reasons: they are subject to the Device Objection, the
Ordinary Motive Objection, the Multiple Motives Objection, the One-Shot Objection and the Unraveling Problem. I then turned to Ainslie's idea of Prescriptive Personal Rules in an effort to solve the Linkage Problem and concluded that there simply can be no such mechanism, and if there were, it would still suffer from all the foregoing objections.

Having abandoned Ainslie's theories of Personal Rules, I turned to three possible variations upon his ideas in an effort to solve the Linkage Problem. The first of these was Nozick's idea of Symbolic Personal Rules, which, I argued ultimately collapses into Principles. I then argued that Principles fail to solve the Linkage Problem in that Principles merely suggest some extended consequences that an agent must consider in making a choice, and in no way explain a boost in motivation in connection with a resolve to take action. Even if engaged intentionally through intentionally reframing a decision, no advance is made since such a strategy is subject to the Device Objection. Finally, I argued that in general there is nothing about the supposition of a symbolic connection between actions and events or states of affairs that *per se* ameliorates the difficulties involved in the Linkage Problem.

I next turned to Bratman's concept of intentions as a possible solution to the Linkage Problem. I concluded that Bratman's account ultimately relies upon secondary mechanisms to explain the stability of intention, and did nothing to advance the explanation of Output Willpower.

Finally, I proposed a creation of my own to solve the Linkage Problem: Will-Based Side Bets. I argued that this mechanism suggested some advance in showing that the added motivation of willpower can flow from the decision to act, rather than from a separate decision to engage a device. However, this mechanism suffers from several
serious defects, including the Proportionality Problem, and a failure to really hit the mark with respect to the Linkage Problem.

Having examined Ainslie’s Personal Rules mechanism and all its apparently available permutations and found them wanting, I now move on to propose my own account of Output Willpower. Two problems immediately loom large: the Linkage Problem and the Special Motivation Requirement. I plan to tackle them head on. Luckily, the lessons learned in surveying these other accounts will hold us in good stead as we move forward.
Chapter Four

A New Theory of Output Willpower

I. Introduction

In this chapter I will propose a new theory of Output Willpower. I will argue that the explanation of Output Willpower lies in an agent's ability to stake his conventional status as a more-or-less powerful and effective agent capable of achieving his aims in the face of obstacles by dint of willpower on the success of actions to which he commits himself. I argue that this feat is achieved by means of performative\(^1\) acts and utterances. This mechanism is functionally isomorphic with the purported operations of Output Willpower on the folk theory. Further, from the standpoint of the acting agent, the motivations involved in this mechanism can pass for the motivations ascribed to Output Willpower under the folk theory. With no other good alternative explanations clearly in view, I argue that this theory represents a substantial advance in the explanation of Output Willpower.

\(^1\) I use the term "performative" to indicate acts and utterances that achieve the kinds of conventional effects described by Austin, Searle and others. I will discuss these effects and set out some of the related terminology in what follows. I deviate from Searle in this usage of the word "performative." Searle would reserve the term "performative" for those "speech acts" that "are performed by way of using a word that names the very type of act being performed." John R., Searle, "Response: Meaning, Intentionality, and Speech Acts," in *John Searle and his Critics*, eds. Ernest Lepore and Robert Van Gulick, (Oxford: Blackwell, 1991), 81-101 at 94-95. Thus, "I order you to leave the room!" is a performative, specifically, an order, in that it makes use of the word "order." Ibid. at 94-95. However, other writers have used the term more generically, as I do. This prevents awkward locutions when one is focusing on conventional effects achieved by means of something other than spoken words, such as physical actions. For instance, one bids at an auction by mere physical gesture. Searle concentrates on construing the act of speaking as a special kind of action that achieves a conventional effect, so he favors the term "speech act." But where one is also interested in actions, it seems awkward and unnecessary to turn actions into a kind of speech that is, as it turns out, a special kind of action. It doesn't seem to make much sense to call such actions "speech acts," and "act acts" is out of the question. The alternative, "illocutionary acts," seems not much better. Thus, like Rappaport, who is also interested in ways of achieving conventional effects other than through the spoken word, I use the term "performative" in this more generic sense. Roy A. Rappaport, *Ritual and Religion in the Making of Humanity*, (Cambridge: Cambridge of University Press, 1999).
II. "Betting" as a Performative

Having found a number proposed mechanisms for Output Willpower unsatisfactory, I now want to go back to the beginning, to the idea of a Side Bet, which underlies, or at least serves as the foundational metaphor, for most of the mechanisms I have discussed. This time, however, I want to look a little more closely at what is actually going on and what the concept of a "bet" and related concepts imply. In particular, I am interested in the performative aspect of betting and related practices, for I believe that therein lies the solution to the problem of Output Willpower.

Under ordinary circumstances, making a bet is, amongst other things, carrying out a "speech act" called a "comissive." Searle, following Austin, defines commissives as "those illocutionary acts whose point is to commit the speaker (. . . in varying degrees) to some future course of action." This definition requires explaining another technical term — "illocutionary act." To keep things simple, I will merely note that Austin describes an illocutionary acts as utterances "such as informing, ordering, warning, undertaking & etc., i.e., utterances which have a certain (conventional) force" such that one performs an act "in saying something," e.g. warning someone, as opposed to an act "of saying something," e.g. speaking the words "I am warning you that . . ." This


\(^3\) Searle, *Expression and Meaning*, at 14.

should be enough to make sense of the basic definition of "commissive." By way of illustration, one might think of a promise as the paradigm case of a commissive.

Rather than plunging further into speech act taxonomy, I will try to breathe a little life into these conceptions by allowing Austin, in his own words, to get across the basic idea of how one does something in saying something:

[If] a person makes an utterance of this sort we should say that he is doing something rather than merely saying something. This may sound a little odd, but the examples I shall give will in fact not be odd at all, and may even seem decidedly dull. Here are three or four. Suppose, for example, that in the course of the marriage ceremony I say, as people will, "I do"—(sc. take this woman to be my lawful wedded wife). Or again, suppose that I tread on your toe and say "I apologize." Or again, suppose I have the champagne in my hand and say "I name this ship the Queen Elizabeth." Or suppose I say "I bet you sixpence it will rain tomorrow." In all these cases it would be absurd to regard the thing I say as a report of the performance of the action which is undoubtedly done—the action of betting, or christening, or apologizing. We should say rather that, in saying what I do, I actually perform the action. When I say "I name this ship the Queen Elizabeth" I do not describe the christening ceremony, I actually perform the christening; and when I say "I do" (sc. take this woman to me my lawful wedded wife), I am not reporting on a marriage, I am indulging in it.  

Thus, the basic idea of doing something through speaking happens in some rather ordinary parts of our lives, but is a striking phenomenon nonetheless, and one that is very important for my explanation of Output Willpower.

I now return to Side Bets. However, I am now interested in the "betting" element of Side Bets from the standpoint of the theory of speech acts. Searle and Vanderveken offer the following analysis of betting:

There is a set of commissives that name joint commitments by both speaker and hearer, a hearer who then becomes a speaker for the purpose of making his contribution to the joint commissive illocution. Two

---

examples are “contract” and “bet.” A contract is a mutual pair of commitments made by two contracting parties. Party A promises to do something for party B in return for which party B promises to do something for party A. The two commitments are not independent; in a genuine contract one is made in return for the making of the other. And this gives rise to the *quid pro quo* element of contracts. In the standard case of betting where one party makes a wager with another party, we have a similar mutuality. In betting on the outcome of a sporting event, the first party promises to pay the second party a sum of money if his team loses; the second party agrees to pay the first party a sum of money if that team wins. ... 

Thus the special feature which distinguishes bets from other sorts of contracts is that bets are joint conditional promises where the antecedent of one participant’s promise is the negation or opposite of the other’s. For example, “I bet $5.00 the Giants will win” is of the form: “If the Giants win I promise to pay you $5.00 and in return if the Giants lose you promise to pay me $5.00.” ... To be a successfully performed speech act a bet has to be accepted. It is not enough for me simply to say “I bet you $5.00 that P.”

This seems a fair description and analysis of the cultural practice of betting (and of contracts). In any event, I will take it as a standard analysis of betting.

The foregoing analysis of betting immediately makes apparent several ways in which a Side Bet does not seem to be a bet at all. Side Bets involve no mutual promises, but only a single (conditional) promise (i.e. the “promise” to “pay” the Side Bet) if, indeed there is any promise at all. Moreover, there need be no one to hear and accept the Side Bet as Ainslie and Becker have construed it. The one-sided nature of the commitment involved in a Side Bet is better analogized to a vow than a bet.

Vows, unlike promises and threats, need not be directed at a hearer. In vowing to do something, I undertake to do it. But I need not undertake to do it either for or against my hearers. I may simply vow to perform better in the future or vow to get revenge on my absent enemies. Vowing

---

furthermore has an additional element of solemnity which is not necessarily present in promising and threatening.\textsuperscript{7}

Better still, is the analogy of a Side Bet to a pledge.

Pledging is much like vowing, only it does not necessarily have the solemnity of vowing. My pledges are undertakings but they need not be undertakings for or against my hearer. When, for example, I pledge allegiance to the flag, I do not in any sense address the flag. The syntax of "pledge" allows it to take both "that" clauses, and, unlike "vow", it also allows it to take nominal direct objects, as in, for example, the sentence "We pledge our lives, our fortunes, and our sacred honor." A pledge is a strong commitment to a future course of action.\textsuperscript{8}

At this point, I think we are on the right track. As a speech act, a Side Bet, and most especially a Private Side Bet, would be a \textit{conditional} pledge taking a nominal direct object. The idea of a bet being a \textit{conditional} promise shows that there really was something significant in the betting metaphor. I know of no name for such a \textit{conditional} pledge, though I think it is a perfectly intelligible concept, and, as I will try to show, one that is familiar. In fact, this seems to be what the idea of "pledging one's honor" is all about, but here I am getting ahead of myself. In any event, some uses of the word "staking" fit this description – again, this will become clearer in what follows. What I want to argue is that in the kind of Side Bets that explain willpower, the agent is staking the status accorded to agents deemed to have a strong will. However, this is all rather abstract, and an example of another social practice involving the staking of one's status would be helpful.

\textsuperscript{7} Ibid. at 193.

\textsuperscript{8} Ibid. at 193-94.
III. Examples of Staking One’s Conventional Status From Honor Cultures

A. Early and Mid-Twentieth Century Japan

In clarifying what I mean by staking one’s status, an examination of some of the cultural practice that involve honor, in one form or another, will be helpful. By way of introduction, I will let Benedict first spell out a relevant variation of the concept honor.

[In Japan, a] good man feels as strongly about insults as he does about the benefits he has received. Either way it is virtuous to repay. He does not separate the two, as we do, and call one aggression and one non-aggression. To him aggression only begins outside the “circle of giri”; so long as one is maintaining giri and clearing one’s name of slurs, one is not guilty of aggression. One is evening scores. “The world tips,” they say, so long as an insult or slur or defeat is not requited or eliminated. A good man must try to get the world back into balance again. It is a human virtue, not an all-too-human vice. Giri to one’s name, and even the way it is linguistically combined in Japan with gratitude and loyalty, has been a Western virtue in certain periods of European history. It flourished mightily in the Renaissance, especially in Italy, and it has much in common with el valor Español in classic Spain and with die Ehre in Germany. Something very like it underlay dueling in Europe a hundred years ago. Wherever this virtue of wiping out stains on one’s honor has been in the ascendant, in Japan or in Western nations, the very core of it has always been that it transcended profit in any material sense. One was virtuous in proportion as one offered up to “honor” one’s possessions, one’s family, and one’s own life. This is a part of its very definition and is the basis of the claim that these countries always put forward that it is a “spiritual” value. It certainly involves them in great material losses and can hardly be justified on a profit-and-loss basis. ⁹

With respect to the concept of shame which is often the counterpart to such concepts of honor, Williams remarks:

The root of shame lies in exposure in a more general sense [than nakedness in fact or as symbol], in being at a disadvantage: in what I shall call, in a very general phrase, a loss of power. The sense of shame is a reaction of the subject to the consciousness of this loss … . ¹⁰

---


This idea of honor should be clear enough.

But can this kind of honor or one's status as honorable be staked in the way I have described? Clearly it can. What is remarkable is that the act of staking it can be made so explicit. Bendict tells us that "[a] borrower may pledge his giri to his name when he asks for a loan; a generation ago it was common to phrase it that 'I agree to be publicly laughed at if I fail to repay this sum.'"\textsuperscript{11} Closer to home, similar ideas are often at work in "pledging one's honor" to fulfill some duty, "staking one's reputation" on one's ability to carry out some task, or "giving one's word" that some statement is true.

\textbf{B. The World of the Norse Sagas}

It may be most useful to understand such concepts of honor and shame as involving merits or demerits, respectively, on a single register of social status.

In a culture of honor as in saga Iceland, shame was talked about as if it were a social state or a concrete thing and only rarely as a feeling. A shame was done to a person by the hostile challenges of another, or one was shamed by one's own failures to maintain standing in the honor group. If feelings were admitted they usually referred to the emotions or physical illnesses that being done a shame would tend to bring about. The shamed person would thus feel rage or outrage, or simply go to bed in despair. Above all shame was a status with an almost juridical aspect.\textsuperscript{12} The Icelanders of the sagas:

... could readily subsume [shame] conceptually into the structures and logic of the key systems of reciprocity: the feud and the gift exchange. A shame is, above all, something that is given or paid over and that needs to be given or paid back and avenged. Thus it is that shame is frequently coupled with verbs for taking vengeance and for paying and requiting.\textsuperscript{13}

\textsuperscript{11} Bendict, \textit{The Chrysanthemum and the Sword} at 151.


\textsuperscript{13} Ibid. at 120.
Further, "[t]he model involves careful scorekeeping, an alternating rhythm of giving and
taking, inflicting and being afflicted."\textsuperscript{14}

Turning to another part of the Norse saga culture, one can see the staking of honor
in the great epic "Beowulf." Upon arriving in the court of Hrothgar and Wealhtheow,
Beowulf addresses a formal boast to Queen Wealhtheow:

\begin{quote}
"I had a fixed purpose when I put to sea.
As I sat in the boat with my band of men,
I meant to perform to the uttermost
what your people wanted or perish in the attempt,
in the fiend's clutches. And I shall fulfill that purpose,
prove myself with a proud deed
or meet my death here in the mead-hall."\textsuperscript{15}
\end{quote}

Thus, Beowulf stakes his honor upon defeating the monster called "Grendel." The saga
writer goes on to describe the following scene later that evening:

\begin{quote}
And the Geat placed complete trust
in his strength of limb and the Lord's favour.
He began to remove his iron breast mail,
took off the helmet and handed his attendant
the patterned sword, a smith's masterpiece,
ordering him to keep the equipment guarded.
And before he bedded down, Beowulf,
that prince of goodness, proudly asserted:
"When it comes to fighting, I count myself
as dangerous any day as Grendel.
So it won't be a cutting edge I'll wield
to mow him down, easily as I might.
He has no idea of the arts of war,
of the shield or sword-play, although he does possess
a wild strength. No weapons, therefore,
for either this night: unarmed he shall face me
if face me he dares. And may the Divine Lord
in His wisdom grant the glory of victory
to whichever side He sees fit."\textsuperscript{16}
\end{quote}


\textsuperscript{15} \textit{Beowulf}, Heaney (trans.) (2000), lines 632-38.

\textsuperscript{16} \textit{Id.}, lines 669-687.
That night, Grendel bursts into the mead-hall and begins his attack on the sleeping Geats and Thanes, but Beowulf grasps the monster from where he had been resting on the floor with his comrades and places Grendel in a crushing armlock. In the midst of this combat Beowulf “recalled his bedtime speech, sprang to his feet, and got a firm hold.”\textsuperscript{17} From this point, Beowulf proceeds to destroy Grendel. It would appear that by his boast, Beowulf not only expressed his prowess, but staked his honor as a formidable warrior. Here the boast may serve to stake Beowulf’s status, though at the same time, it may have served other purposes such as impressing Wealththeow, for it “pleased the lady well.”\textsuperscript{18}

This pledging of one’s conventional status by means of performative utterance such that it is forfeit in some degree or other if one fails in the task against which one has pledged it, I will call “\textit{Status Staking}.” At this point we have found a means to “arbitrarily” place side stakes as an added motivation for one to carry out a challenging or difficult action, just as Ainslie had hoped to do with Prescriptive Personal Rules. However, the pledge is without real significance if it is without binding force. That is the concern to which I next turn.

\textbf{IV. The Self-Enforcing Nature of Status Staking}

One fact of fundamental importance about the kind of Status Staking I have been discussing is that it is \textit{self-enforcing}. Returning to the Japanese borrower of former times who “pledge[s] his giri to his name when he asks for a loan” and “agree[s] to be publicly laughed at,” Benedict tells us that:

\begin{quote}
If he failed [to timely repay the debt], he was not literally made a laughingstock; there were no public pillories in Japan. But when the New
\end{quote}

\textsuperscript{17} \textit{Id.}, lines 757-59.

\textsuperscript{18} \textit{Id.}, lines 639-641.
Year came around, the date on which debts must be paid off, the insolvent debtor might commit suicide to ‘clear his name.’ New Year’s Eve still has its crop of suicides who have taken this means to redeem their reputations.\(^\text{19}\)

Hence, neither the debtor, nor his creditor, nor anyone else need do anything to inflict damage upon the giri to his name. That damage was done when the debtor failed to repay his debt by virtue of the constitutive rules or cultural practices and understandings that are constitutive of the honor understood as giri to one’s name. By committing suicide, he is trying to redeem honor that has already been lost by virtue of his failure to carry out the action on which he staked his honor.

But how, precisely, does this self-enforcing aspect of Status Staking work? Again, there are two components, the conventionally ascribed status, and the performative act or utterance that creates, confers or modifies it. For instance, there is the status of being a black belt in Tae Kwon Do, and various performances that confer the status of black belt, or of moving from a first- to a second-degree black belt. Examples of this kind abound. Two old philosophical favorites are the status of being married and the status of being knighted. In each case, the status is transparently socially constructed by constitutive rules, and conferred by following the prescribed procedure. Nonetheless, whether or not one is married or knighted is regarded as a matter of fact, and such a fact, such a feature of one’s life, may be something highly valued by the affected agent.

It has long been remarked that the efficacy of performatives needn’t be dependent upon the subjective mental state of the agent whom they bind. Consider that paradigmatic commissive, the promise. As Austin noted, a promise, even if insincere, is

\(^\text{19}\) Benedict, *The Chrysanthemum and the Sword*, at 151.
a promise nonetheless, as we still hold the promissor at fault if he fails to keep it.\(^{20}\)

Searle writes more generally that:

\[\text{[W]hen one enters an institutional activity by invoking the rules of the institution one necessarily commits oneself in such and such ways, regardless of whether one approves or disapproves of the institution. In the case of linguistic institutions like promising (or statement making), the serious utterance of the words commits one in ways which are determined by the meaning of the words. In first-person utterances, the utterance is the undertaking of the obligation. In certain third-person utterances, the utterance is a report of the obligation undertaken.}\(^{21}\)

Rappaport has relied on the distinction between the subjective state of belief and what he refers to as “acceptance” in his complex and wide-ranging theory concerning the ways in which rituals’ performative acts and utterances create obligations and bring about various conventional states of affairs. The distinction is crucial for Rappaport, for he argues that in at least some cases, participation implies acceptance, thus enabling ritual to bring about these conventional states of affairs without being wholly reliant upon the subjective states of belief of all participants.\(^{22}\)

Yet all this shows is that, in formal terms, if one pledges one’s honor to complete a task and fails, one loses the implicated conventional status. This still shouldn’t provide for meaningful enforcement of the pledge if the agent is indifferent to the conventionally

\(^{20}\) Austin, “Performativie Utterances,” at 235-36.


ascribed status. Perhaps one has gained a self-enforcing pledge at the cost of making the stakes forfeited purely formal.

This objection is good as far as it goes. However, I think that we should not overestimate our ability to choose which conventional statuses we care about. We measure ourselves by any number of indicia which, while they may be valued by the culture at large, we as individuals often “officially” renounce as meaningless: beauty, weight, height, ethnicity, race, family background, economic status, “manliness,” “femininity,” etc. The list could go on and on. Miller has made something of a study of how honor and shame, those fundamental categories of status, still retain vitality despite their official abandonment.

Honor is not dead with us. It has hidden its face, moved to the back regions of consciousness, been kicked out of the public discourse regarding individuals ...; it can no longer be offered as a justification for action in many settings where once it constituted the only legitimate motive. But in spite of its back-alley existence, honor still looms large in many areas of our social life, especially those, I would bet, that occupy most of our psychic energy.  

Thus, we may still care about status assessments we nominally disclaim and regard as pointless or arbitrary. Once we participate in such an institution of status making and status breaking, it may still exert for us the old pull, the affective ties that bind. In the light of recent work in moral philosophy by Williams and others showing how deeply we care about aspects of ourselves as well as effects of our actions which may be, from a moral standpoint a matter of arbitrary luck, these conclusions should be less surprising than they once might have been.  

---

23 Miller, *Humiliation.*

Against this background, it is helpful to revisit that shop-worn truth of the anthropologist, that the values and status ranking that matter to individuals may often be dependent upon and largely determined by cultural context. Again, Benedict:

Giri to one’s name and all the hostility and watchful waiting that accompany it in any culture, however, is not a virtue that is characteristic of the Asiatic mainland. It is not, as the phrase goes, Oriental. The Chinese do not have it, nor the Siamese, nor the Indians. The Chinese regard all such sensitivity to insults and aspersions as a trait of “small” people—morally small. It is no part of their idea of nobility as it is in Japan. Violence which is wrong when a man starts it out of the blue does not become right in Chinese ethics when a man indulges in it to requite an insult. They think it is rather ridiculous to be so sensitive. Nor do they react to a slur by resolving by all that is good and great to prove the aspersion baseless. The Siamese have no place at all for this kind of sensitivity to insult. Like the Chinese they set store by making their detractor ridiculous but they do not imagine that their honor has been impugned. They say “The best way to show an opponent up for a brute is to give in to him.”

Or, to put the shoe on the other foot:
The idea that the pursuit of happiness is a serious goal of life [as it is for Americans] is to [the Japanese] an amazing and immoral doctrine. Happiness is a relaxation in which one indulges when one can, but to dignify it as something by which the State and family should be judged is quite unthinkable. The fact that a man often suffers intensely in living up to his obligations of chu and ko and giri is no more than they expect. It makes a hard life but they are prepared for that. They constantly give up pleasures which they consider in no way evil. That requires strength of will. But such strength is the most admired virtue in Japan.

While Benedict’s account may now be dated, it nonetheless makes the relevant point. As a matter of fact, people do care about various kind of conventional statuses. The whys and wherefores of such caring is likely to be made most sense of by looking inside the network of connected social and communicative practices that constitute the relevant culture.


26 Ibid. 192.
In sum, if as a matter of fact, one is under the sway of a certain culturally designated status, its achievement or maintenance can be highly motivating, and obviously if one is not impacted by it, it does little to motivate behavior. Again, the example of honor.

To the extent that a person’s very social being is dependent upon one’s being honorable, one must palpably feel the loss of honor, that is, shame. The person who does not subscribe to the norms of honor will not feel shame for having violated them even if real third parties try to make him or her feel so. This invulnerability is simply an aspect of the social quality of shame. Shame requires membership in a society, a community of people sharing norms of right action and caring deeply about what others in the community think of them.²⁷

Importantly, Williams in his study of Classical Greece notes that shame, at least at a certain level of development, does not require an actual witness to one’s shaming; no one need, to revert to the example of the Japanese borrower, actually laugh one, “the idealized other will do.”²⁸ Miller confirms Williams’ observation regarding the internalized nature of shaming in historical practice:

Being shamed … did not necessarily require an observing audience. One needed only to be committed in a serious way to the values and standards of the community in which one claimed membership to feel shame (and to be shamed) for not measuring up to those standards or adhering to those values.²⁹

Thus, if one is in their sway, such conventionally conferred statuses can have a powerful emotional bite, and this emotional impact need not be brought about through the actions of any second party.

²⁷ Miller, Humiliation, at 118.

²⁸ Williams, Shame and Necessity, at 221.

²⁹ Miller, Humiliation, at 134.
Beyond immediate emotional effects, the conferral of a conventional status or ranking may have further material effects. Again, the example of honor.

Honor, *like strength of will*,\(^{30}\) may imply the attainability of other, further, rewards. Honor translated into practical advantage. It could be practicality itself. It meant good marriages for oneself and kin; it meant active involvement in a number of exchange cycles with other people of honor, whether the exchanges were of gifts or of insult and injury. Honor, above all, meant relations of reciprocity with other honorable people. The life of honor could thus accommodate peace and peaceful resolution. ... Any waiver of blood revenge could only be honorably made if one was able to indicate one's future inviolability at the same time. It could be said that honor is the ability to make others believe that you will indeed be tough the next time, in spite of present discomfiture.\(^{31}\)

This passage makes evident some of the further practical effects that may flow from the ascription of a conventional status.

At this point, I have shown that Status Staking is a kind of self-enforcing conditional pledge. Further, within the right cultural context, a given status ascription pledgable through Status Staking may be of great emotional or practical significance to the agent, such that he is highly motivated to attain or maintain it. Thus, once staked, even through a more or less arbitrarily constructed pledge, such a status ascription is both highly motivating and irrevocably staked on the agent's performance.\(^{32}\)

**V. Further Refinements of the Mechanism of Status Staking**

**A. Introduction**

Ultimately, I want to argue that agents stake their status as strong-willed persons when they resolve to take action. However, Status Staking can occur in a variety of

---

\(^{30}\) This analogy between honor and strength of will is one I will pursue in what follows.

\(^{31}\) Miller, *Bloodtaking and Peacemaking*, at 303 (emphasis added).

\(^{32}\) It is worth noting that since this mechanism operates on the basis of status being conferred, rather than merely evidenced, there is no question of the issues implicit in Newcombe's Problem serving as a basis for objections to the "rationality" of Status Staking.
ways, and I have not yet given any examples of the right sort of model of this process. I will build my case by giving examples that demonstrate the requisite features of the kind of status staking with which I am concerned. Along the way I will show that these examples comport with the theory of performatives.

B. Ways of Staking One's Status

In trying to construct the appropriate model of staking, I am interested in cases where the agent's affirmative acts put his own status at stake. I want to contrast these kinds of cases with those in which honor is put at stake merely because one has been acted upon, rather than because one has acted. In the latter cases it is the other who puts one's honor at stake. The classic example of this occurs in the revenge cycle, or balanced exchange of honor and shame that occurs in the Icelandic sagas as described by Miller. Honor is put at stake when one is shamed by another, and the shamed individual need do nothing to have his honor so staked other than receive the shaming injury or insult. The shamed individual becomes active only in trying to redeem his honor by avenging himself.

I have already provided examples of cases where an agent takes affirmative steps to put his status at stake in honor cultures, but I now want to classify them a little more carefully. In the first kind of case, the agent more or less explicitly announces that he is staking his honor. Here, the agent may state that he "pledges his honor" to accomplishing some aim. Only slightly less explicit is the agreement of the Japanese borrower to be metaphorically made a laughingstock if he fails to pay his debt on time.

In the second class of cases, status is staked intentionally, but without announcing that that is what one is doing. Beowulf's formal boast may be an example of this. That
one does not use the words “I pledge my honor” poses no particular difficulty. As Austin remarked, we can certainly carry out the performative of insulting someone with our words even though we do not say the words “I insult you.” In fact, such a locution would be extremely odd and would lack any force as an insult, though in some unusual circumstances it might serve make one who was unable to understand the intent of some previous utterance grasp that it was intended as an insult.

For the third kind of case, I will introduce another example. Veyne writes of an honor-based expressive act of bestowing gifts on the community in Roman and Greek antiquity.

Eugertism means the fact that communities ... expected the rich to contribute from their wealth to the public expenses, and this expectation was not disappointed: the rich contributed, indeed, spontaneously and willingly. Their expenditure on behalf of the community was directed above all to the entertainments in the circus or the arena, and, more broadly, to public pleasures (banquets) and the construction of public buildings—in short, to pleasures and public works ... .

One of the motives sometimes operative in such giving is associated with the concept of “munificence” or “magnificence.” Such eugertism:

...could be connected with themes of which it constitutes a historical modification. In the first place, patronage, which Thorstein Veblen satirically equated with ostentation, but which is, rather, the effect of a tendency possessed by individuals and groups to actualize their potentialities, together with a tendency to express their superiorities, even for their own satisfaction, without any audience being present.

The motivating desire might be the

---

33 Austin, How to Do Things with Words, at 65.


35 Ibid. at 70.
desire to display one's wealth, to express one's personality, to put oneself forward so as to stand out from the people (especially when the rich man was aiming at a career as a political orator), desire to leave one's name on a "work", the spirit of competition.36

At the same time, such displays of wealth could increase the prestige of the giver, heighten his honor. In fact, eugertism could also be motivated by a desire for honors, "[s]pending so as to be honoured."37 Of course, such a motive need not always be present, and the honors could be accorded nonetheless.

This example is rather complex, but for now I want to focus on one feature of eugertism. According to Veyne, a notable might undertake to construct some public building in order to display his wealth and express his personality. This would be primarily a demonstration of abundance and power. The notable would need no further reason to undertake the project. Specifically, it is not necessarily the case that he would aim to enhance his status, but rather that he might merely wish to display that power, abundance, and generosity of spirit which he felt quite assured that he already possessed. Nonetheless, upon announcing such an undertaking, or beginning it and proceeding far enough that his intentions become known, he also stakes his status upon carrying it off successfully. Even the self-assured notable that I have described may suffer damage to his status if he lacks sufficient wealth to complete the project, or if the workmanship is of poor quality, or if the design is an aesthetic failure. In undertaking the project, the notable has staked his honor—even if that was not an effect he intentionally sought to create. Moreover, it is difficult to imagine, within the culture of honor, how he could immunize his actions from such implications if he were to want to do so.

36 Ibid. at 78.
37 Ibid. at 78.
The idea that an act or utterance may have dual performative functions has already been recognized within Searle’s theory of speech acts.

Consider Suzanne, a prototypical liberal, who says, “I protest the way you are treating the poor.” Her protest is certainly an expressive speech act in that it reflects back on the treatment of the poor and expresses her attitude toward that treatment. However, as Searle points out, a protest can be, and we can assume in Suzanne’s case is, a recommendation for change. So her protest is an expressive directive or, if you like, a directive expressive.38

We can imagine contemporary examples of such dual performatives involving actions rather than words. For instance, an opera diva may sing an aria primarily for the purpose of displaying either her own abilities or perhaps to display the beauty of the aria itself. She expresses something about either the aria or about herself. Nonetheless, in the same performance she may be staking her reputation with critics and her audience upon the skill with which she performs the aria.

It poses no particular problem that a commissive is established by action rather than speech. Austin noted that one can, for instance, “say,” in effect, “I am protesting” by chaining oneself to park railings.39 Examples of this kind abound, even with respect to commissives. For instance, a Maring man obliges himself to lend military assistance to another tribe by dancing at its kaiko celebration40 and one bids in an auction (thereby binding himself to pay the amount bid if his bid is accepted) by any of a variety of physical movements. With respect to staking honor, specifically, one can imagine two angry motorists stepping out of their cars after an accident and facing each other down in

38 Fotion, John Searle, at 58.
39 Austin, How to Do Things with Words, at 64.
40 Rappaport, Ritual and Religion, at 113-16.
preparation for coming to blows. The first motorist may swagger toward the second in a way that expresses his physical prowess or lack of fear, but that at the same time stakes “face” or “honor” on his ability to live up to the representations implicit in these expressions. In fact, by his “macho” display the first motorist may be making a triune performative in that he carries out actions that are expressive, stake or pledge honor, and implicitly assert that he has what it takes to back up the bravado he expresses.

Nor is it any problem to suppose that Status Staking may be effected by “inner speech,” to use James’ expression. Other familiar performatives are expressed in this way. A Roman Catholic priest may bless or absolve the parishioner without a spoken word, and in principle without even an overt gesture. “Inner speech” may be sufficient. Nonetheless, the penitent is, as far as the church is concerned, absolved, and the devout are blessed. We may disagree with the idea that these acts render any metaphysical benefit to the parishioner, but as a matter of fact the recipient of these offices is absolved or blessed by the priest. His conventional state vis-à-vis the institutional structure of the church has been altered.

Further, even without clearly identifiable inner speech the very formation or maintenance of an intention gives rise to a kind of commitment to action.

Why must my making up my mind to do so-and-so, which entails that I become certain that I will try to do this, be distinguished from the prediction that I will in fact do it. The principle ground of this distinction seems to be a prior distinction, taken for granted in the use of ordinary language, between that which I myself do at will and that which happens to me. An attempt is by definition something that I do at will. Failure in the attempt is something that will happen to me contrary to my intention whether the failure is culpable or not. Suppose that I decide to try to do something and at the same time decide, on the evidence of past performances, that I shall probably fail. It seems that I cannot, in the normal use of the words, call the first decision a prediction and that I must call the second decision a belief, which, if it were expressed, would be
called a prediction. If I call an expression of the first decision a prediction, I am thereby implying that an attempt will not be made of my own free will. If I deny that an announcement of the second decision is a prediction, I thereby imply that I am going to make myself fail, that is, that I will not really try. I may take steps which, from the point of view of an observer, may seem to be an attempt to do something, but they cannot be counted as a real attempt, if I have never had the least intention of achieving what I am aiming at. 41

Thus, without making any metaphysical or even scientific claims about intention, we can say that in our ordinary use of the concept, we take intentional decisions to take an action to imply that we will attempt the action.

Where one fails to carry off the attempt successfully, the fault may sometimes be put on the agent for not living up to the intention, but he will typically not be blamed for having made an erroneous prediction about his future conduct. For instance, if Pete suggests to me that he will meet me at the library tomorrow, and he fails to do so because at that time he prefers to do something else, I will not criticize Pete for making an erroneous prediction about the likely state of his personal utility function or of his motivational system at the time he would have had to take action in order to meet me at the library. Rather, I will fault him for failing to live up to his stated intention. That is, I will not fault Pete with a failure of knowledge, but rather with a failure of action—a certain fickleness or weakness in failing to follow through with plans. I do not treat his statements of intention about future action as predictions about what he is likely to find himself doing; rather, I treat them as statements about actions that he is committing himself to taking.

This is different from a promise—for he needn’t have promised me that he would be there, even implicitly, for me to have this reaction. Suppose, for instance, that Pete indicates that he will go to library in order to do research which he needs to have done by Monday in order to properly deliver a lecture. If he fails to do so simply because he prefers not to at the time, I will fault him with this same weakness or fickleness. Pete has broken no promise, for there is no one to whom such a promise could have been made. Nor need we suppose that a vow was made, silently or otherwise. Nonetheless Pete had, in a sense, committed to the action of going to the library, and certain status attributions ride on his ability to follow through on that commitment. Similarly, I would treat my own failure to live up to an intention to go to library simply because at the moment of action I preferred not to as a failure of action—assuming that all the reasons for my so doing, including the consequences for me and others, and more specifically, the desires that I will have tomorrow about these events, remain the same. From this it follows that Status Staking may occur merely by the formation of an intention to the extent that some valuable self-conception is associated with following through, or being capable of following through, on one’s intentions.\footnote{In this I may be parting company with Searle who simply announces without any real justification or analysis that “[i]ntending is never a speech act; expressing an intention usually, but not always, is.” John R. Searle, \textit{Expression and Meaning: Studies in the Theory of Speech Acts}, (Cambridge, Cambridge University Press, 1979), 9.}

\section{C. Staking by Degrees}

Another important fact about Status Staking is that it may be done by degrees. As already noted, Miller tells us that for the Icelanders, “[a] shame is, above all, something that is given or paid over and that needs to be given or paid back and avenged”\footnote{Miller, \textit{Humiliation}, at 120.} and this
model "involves careful scorekeeping." Benedict quotes a Japanese author who writes that:

In revenge there is something that satisfies one's sense of justice. Our sense of revenge is as exact as our mathematical faculty and until both terms of the equation are satisfied we cannot get over the sense of something left undone.

Thus, one's honor may be impugned in varying degrees depending upon the gravity of the dishonor one suffers.

Just as the sense honor can be staked in varying degrees depending on the slight of another, so too can it be staked depending upon the actions one takes which constitute the placing of the stake. This variation in degree can arise in at least two ways. Normally, one does not pledge, for example, one-fourth of his honor. Still, the amount pledged may be implied (rather than stated) by other features of the pledging performative. Consider a young man, Sam, asking a woman, Sally, out on a date. If Sam expresses a great deal of desire in his request and Sally rejects his invitation, Sam may lose more "face" than if the invitation was merely casual. Thus, the amount of "face" at stake may be roughly proportional to the strength of the expressed desire. Second, the degree to which status is staked may be a function of the circumstances in which the pledge is made. Consider the difference between Sam asking Sally out for coffee after work, and asking her to a concert for which he has already purchased costly tickets. A refusal will certainly result in a higher "cost" to Sam in the latter case (ticket prices aside).

---

44 Miller, Bloodtaking and Peacemaking, at 182.

45 Benedict, The Chrysanthemum and the Sword, at 161 (quoting Nitobe (1900), at 83).
Other cases follow a similar pattern. Generally, there is more at stake for the person who says to himself “I will finish this job if it is the last thing I do!” than there is for the person who says, “I am going to give it the old college try.” As for context, if one looks at the daunting prospects before him and nonetheless says, “I will do it!” it seems there is generally more required of him and more at stake in terms of his own self-evaluation than if he were taking on some rather trivial chore such that, even if he fails to carry it out, he is confident that he could do anyway “if he really set his mind to it.” Thus, with respect to the latter class of cases, one’s resolve to undertake action in the face of difficulties may implicitly or explicitly, depending upon how it is expressed, signal a level of commitment which corresponds to the size of the stake in terms of degree to which one’s status is staked on carrying out the action.

How much is at stake may also be modulated during the course of carrying out the action to which one has committed oneself. Consider again the two motorists facing each other down after the accident. The more one expresses his willingness to fight, the more his pride or honor may be staked on his not backing down. These expressions may be in words, inner speech, bodily actions, or qualities of actions such as the manner in which one driver approaches the other. These may involve eye contact, swagger, verbal communication, hand gestures and so on, each of these serving to express physical prowess, fearlessness, or malevolent intent. Any of these modes of action may effectively trigger the commitment involved in staking so long as they are normally understood as doing so in the relevant community and the individual undertakes them knowingly. Hence, the whole performance of facing each other down may involve any number of component performances that are elements of the larger episode. Each of
these component performances may in turn be modulated according to circumstances. Throughout each of these sub-performances, and through their modulation, the amount of "face" staked may be increased and, at the same time, be partially redeemed or at least preserved as one escalates the level of his commitment by showing that he is not intimidated.

D. Conclusion

With this, I conclude my discussion of some of the refinements of the Status Staking mechanism. I have shown how status staking may occur without a direct expression that that is what one is doing, through dual or even triune performatives. I have shown that the relevant performative expressions may be in words, deeds, thoughts ("inner speech") and even in the formation and maintenance of an intention. I have also shown that status may be staked to varying degrees vis-à-vis the attainment of a given objective, and that this variance can be achieved either through using different or modified performatives or through uttering or doing the same performative in different circumstances. Having set out the Status Staking mechanism in some detail, I now move on to discuss that particular status that will be "plugged into" this general mechanism in an effort to explain Output Willpower.

VI. The Status of the Strong-Willed

Ultimately, I hope to show that what is at stake in the pledge operative in Output Willpower is the agent's status as a strong-willed agent. However, a good deal of work must be done before that can be established. The first step is to argue that being regarded as strong-willed is a conventional status agents care about.
As a preliminary matter, we have assumed that under BMA Output Willpower construed literally, in its magical or romantic sense, is simply an impossibility. So there can be no matter of fact that one has a strong will in this sense. One might say that there is no metaphysical fact to uncover. However, attributions of strength of will are made according to culturally regulated practices, and it is these attributions with which we are concerned. It is these attributions that confer a valuable social status upon the agent. It is for this reason that the analogy to honor is particularly apt.

That such status attributions may be bound up with false theories about the world is not an argument against a status being conferred upon an agent nor against the relevant status being highly prized. We can put to the side for the moment whether attributions of willpower are dependent upon a fundamental mistake about the facts, like attributions of witchcraft, or whether any such false theories are merely incidental. An example of the latter is the theory of divine right. Such a theory may or may not be a view held by the king’s subjects, but the fact of the matter as to whether or not they hold such a view is irrelevant as to the determination of whether an individual is in fact king. The latter status is conferred by means of an appropriately performed coronation under the right background circumstances. Of course, a belief in the divine right of kings may help to motivate the maintenance of the institution of the monarchy, but that is a separate matter.

It seems clear that, as a matter of fact, agents care about whether they are deemed strong-willed or weak-willed. Willpower has something like the status associations of an executive virtue like courage or self-control: it enables one to act in accordance with one’s endorsed aims where he would otherwise be unable. These executive traits, in turn, become valued in and of themselves, at least to the extent that they are not put to evil
purposes. Even where they are, the agent may take a kind of pride in these qualities of character. As a simple matter of fact, one may be admired for being a daring thief, a fearless warrior, or an adventurous explorer without any pretension of there being a connection between these qualities of character and conventional moral virtues. Further, as a matter of psychological fact, such a person may look upon himself as possessing any one of these respective characteristics with pride. Whether this should be so is another matter entirely.

Of course, I do not want to claim that having strength of will is a virtue, executive or otherwise. To do so would raise issues of Output Willpower's connection with ethical conduct or rational prudence either historically or according to our best contemporary understanding of these matters. I already indicated at the outset that I do not think that there is any such strong connection, and I think it is wasted effort for my purposes to argue whether Output Willpower can or cannot be assimilated to the model of an executive virtue. Perhaps most importantly, it would unnecessarily embroil me in debates about the role virtues play in motivation. I have my own story to tell about motivation, and though it may at times be useful to draw upon the philosophical work done in connection with the theory of virtues, this story can and should be told independently of an account of the virtues.

Perhaps I am on safer ground in drawing the analogy between strength of will and physical strength. On the folk theory, both make a wider range of actions possible in an obvious and immediate way. Both are highly valued for this reason. Of course, how much either of these traits is valued is largely relative to a given community, and this in turn may depend upon either contemporary or historical circumstances requiring exercise
of the respective form of strength to achieve purposes regarded as important within that community. If one is part of a warrior band, for instance, physical strength will be highly prized. In many parts of the modern West, the ability to delay gratification and follow through on the course one has set for oneself is, and has been, highly prized. This is not surprising, given that one's success in such societies is likely to depend in large measure on one's ability to follow through on one's intended action, even where that requires delaying gratification. The alternative is having one's plans undone through succumbing to ready distractions, more easily obtained specious rewards, emotional indulgences, or the simple lack of the motivation to take appropriate action.\footnote{Ainslie remarks on the importance of willpower in the circumstances of relative abundance in which persons in the modern West find themselves. George Ainslie, \textit{Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person}, (Cambridge: Cambridge University Press, 1992), 154. For a lament for the alleged decline in such personal self-determination since the Victorian Era, see Gertrude Himmelfarb, \textit{The De-Moralization of Society}, (New York: Vintage, 1994), 221-258.} Having the requisite enabling strengths in the relevant community often occasions pride, whereas lacking them often fosters shame.\footnote{Instructive in this regard is what Williams has to say about the connection between shame and accidental features of oneself or one's history, that is, features one has not freely chosen in the way that Williams identifies as what "morality" would require before a "moral" assessment is warranted. Williams (1993).} (This is, perhaps, part of the point of Miller's analogy between the enabling features of being regarded as strong willed and being regarded as a person of honor, respectively, as quoted in the last section.) I take these as more or less commonly accepted facts.

Finally, being strong-willed is particularly desirable because it is taken to be closely connected to autonomy. One is able to carry out one's choices because one wills it—one freely chooses to carry them out. This ideal is not so well realized by, for instance, Ulysses binding himself to the mast, or a person having his jaw wired shut to prevent overeating, because each employs an external device in order to achieve their
respective aims. Nor is our concept of autonomy so clearly implicated by the more subtle mechanisms or devices that I have described thus far in this dissertation. Thus, willpower is closely associated with the ability to achieve one’s aims autonomously in the face of distraction, temptation and exhaustion merely because one has resolved to do so. Thus, for a variety of reasons, being strong-willed is a status of real significance to the agent.

VII. Functional Isomorphism

We are now in a position to plug the status of being strong-willed into the Status Staking technique or mechanism. In so doing, we will have achieved a significant portion of the explanation of Output Willpower in that the result is a mechanism which can have the same application and the same versatility that is associated with the functioning of Output Willpower on the folk theory.

On the folk theory, willpower is engaged when one resolves to take action, perhaps accompanied by “inner speech” affirming that one will or must carry out the action that will ultimately be motivated by willpower. I have argued that such an affirmation may stake a status such as that involving strength and power to which honor is accorded. Similarly, it seems that such an affirmation could stake one’s status as being strong-willed and capable of carrying out one’s resolutions to action under the folk theory of willpower. It may be that to form an intention, i.e. resolve to undertake an action, in the face of known temptations that would lead one astray is all that is needed to stake one’s status in terms of self-attributions of inner strength. One sees the temptations, or the difficulty or the danger involved in a course of action, and resolves that one will do it anyway. In the right context, one has staked one’s status of being capable of firm resolve on carrying out the task.
As a result of the staking that arises as a consequence of one's resolution to take action, a new motivation arises for succeeding in that action—i.e. not losing that which is staked on the success of the action, and perhaps seeking to increase it. This motivation operates causally such that it may give the agent the motivational power to carry out the task that he otherwise would have been unable to perform and, like willpower, it arises as a result of one's resolution to act. This same mechanism or technique may be triggered regardless of the sort of action one resolves to undertake. Thus the corresponding motivation may be applied to virtually any goal, whether morally good or bad, prudent or imprudent. The status attribution of strength or power that elicits this motivation bears no necessary connection with attributes of the moral or rational worth of a given course of conduct.

It is also the case that the manner in which the resolution to act is expressed, or the context in which it is expressed, can cause variation in the strength of this motivation. A resolution to take action made in the face of a known and substantial obstacle, or made more emphatically, may tend to create a greater stake in success and thereby increase the motivation to succeed just as on the folk theory one serious about undertaking an action calls upon greater resources of will to undertake the task at hand. The idea of the variability of effort is often expressed in terms of such phrases as "not really trying," "giving it a shot," "really trying," "trying hard" or "giving one's all." Each of these expressions may be applied to the same primary aim or objective that has, in each instance, the same intrinsic reward value (i.e. reward value considered without the added side stake).
If these conclusions are correct, the mechanism I have hypothesized exhibits a kind of functional isomorphism with Output Willpower as that power or faculty is described within the folk theory.

VIII. The View From Inside

A. Introduction

Having constructed a mechanism that functions isomorphically with Output Willpower, the next question is whether it will pass for Output Willpower from the standpoint of the agent engaging the mechanism. I will argue that it can for two very different kinds of reasons. First, I want to argue that agents can and do accept the mechanism I have just described, albeit understood intuitively rather than formally, as Output Willpower without any contradiction or irrationality. Second, agents may engage this mechanism without being aware of its nature, and in fact, be deceived by an overlay of theory about its functioning that is associated with folk accounts of willpower, thereby leading them to attribute to Output Willpower some of the more “magical” properties ascribed to it under the folk theory.

B. Accepting Status Staking on Its Own Terms

The first argument, that agents may accept Status Staking as the mechanism of Output Willpower with more or less open eyes, is a short one, but one that I hope has been implicit in much of what I have already said. First, agents regard actions and dispositions of character as worthy of praise or condemnation regardless of whether or not they spring from the deliberate and intentional choice of the agent. I have tried to argue that this is so in honor cultures in terms of the strength one exhibits, whether as a warrior, benefactor, or simply as a man of honor. In no case does the status accorded to
one exhibiting such qualities of character hinge upon his having achieved the underlying
capability as a result of free choice in some metaphysically deep sense. Williams
provides the following illuminating example from Greek tragedy:

Slighted by the award of Achilles’ arms to Odysseus, [Ajax] plans to kill
the leaders of the army. To prevent this, Athena makes him mad.
Thinking that he is killing Odysseus and the others, he slaughters the
army’s flock of sheep and cattle .... He is first seen unconscious, covered
in blood among the mangled beasts ... .

Ajax then wakes up and shows that he has recovered his mind. There is a
passionate lyric outburst of despair and, above all, shame: he has made
himself, apart from anything else, utterly absurd. It becomes increasingly
clear to him that he can only kill himself. He knows that he cannot change
his ethos, his character, and he knows that after what he has done, this
grotesque humiliation, he cannot live the only kind of life his ethos
demands. ... Being what he is, he could not live as the man who had
done these things; it would be merely impossible, in virtue of the relations
between what he expects of the world and what the world expects of a
man who expects that of it.48

Such a reaction is certainly intelligible, and, as I have tried to make plain,
common, though this is an extreme case. Williams goes further in arguing that we are
mistaken if we think that in matters of ultimate import, i.e. moral matters, praise and
blame hinges upon the relevant action having been undertaken by dint of a
metaphysically free choice.49 I need not take the argument so far. I merely want to assert
that such attributions of praiseworthy, on the one hand, and blameworthy,
contemptibleness, or worthlessness, on the other, are commonly made with reference to
such “contingent” or “accidental” features of character or behavior, and that these
attributions are intelligible and probably practically unavoidable.

48 Miller, Humiliation, at 72-73.

49 Id., Chapter Three.
Granting these conclusions, it is a short step to being able to take Status Staking as Output Willpower with open eyes. For one may regard the threat of humiliation felt at the failure to carry out one's resolve, or the promise of esteem to be felt as succeeding at the same, to be natural and appropriate emotional responses without need of any metaphysical underpinnings in a causally independent, magical concept of willpower. One may take it as obvious that these are the motives of action in the cases at issue, or, at least, one might take it as such if the point were pressed. The struggle to maintain honor, for instance, may be seen to involve a kind inner strength, even though it is obviously motivated by the apparent stake one has in preserving it, i.e. avoiding humiliation rather than by virtue of some metaphysically potent faculty. Again, Benedict's remark (though now with a different emphasis):

The fact that a man often suffers intensely in living up to his obligations of chu and ko and giri is not more than they expect. It makes a hard life but they are prepared for that. They constantly give up pleasures which they consider in no way evil. That requires strength of will. But such strength is the most admired virtue in Japan.\(^{50}\)

Thus, defending one's status, including one's status as strong-willed, may itself be one of the criteria of being strong-willed.

This is connected with the next point, that within such a system these motivations may seem not only natural but also appropriate in the sense that those who lack them are in some way deficient. This idea is often expressed with such words as "You should be ashamed!" or, more archaically, "Have you no shame?" addressed to one who doesn't seem to feel or be motivated by the expected sense of humiliation or the fear of incurring

\(^{50}\) Benedict, The Chrysanthemum and the Sword, at 192.
it. From a standpoint within such a cultural system, one ought to have these prescribed
reactions to having acted in the relevant ways.

I think, then, that there is no contradiction and no irrationality in an agent openly
acknowledging Status Staking as the mechanism of Output Willpower, nor in not feeling
any great need to "deconstruct" that mechanism intellectually, or to try to purge it from
one's motivational repertoire—whatever that would mean. Even if one takes the extreme
position that all such reactions ascribing merit and demerit to behavior are irrational
unless the evaluation attaches to metaphysically free actions, it remains a fact that there is
at least no unusual irrationality attaching to the status ascriptions connected with
attributions of willpower on this theory. Further, given the psychologically trenchant
nature of these status attributions, I see no reason why even one who "officially" held a
metaphysical theory of willpower would suddenly lose his willpower upon learning that it
was merely Status Staking, and that he was in fact mistaken as to what is going on when
he exercises his willpower. This last claim will be further supported by the following
discussion as to how it can be that one is deceived about the operations of Output
Willpower.

C. The Metaphysical View of the Will

I have argued that for some the folk theory may imply that willpower is a faculty
or a power that operates outside of the laws of nature, that it is a force that has no causal
antecedents and can vary in strength independently of causes and consequences. How
could such a view be compatible with the position that Status Staking is willpower? I
have tried to make clear from the start that from the standpoint of science and the BMA,
nothing could literally answer to this description. So the question really is: How can the
operations of Status Staking appear to agents operating within the folk theory of willpower to evidence a metaphysical faculty of willpower?

The first step down the road to the explanation involves the recognition that people ordinarily expect to be capable of exercising willpower, therefore they are not necessarily focused upon what is at risk in their exercise of it. Again, an analogy to honor. In some cultures, those in a position of honor do not feel themselves constantly under threat of losing honor because they feel secure in those qualities or power on which the ascription of honor is based. They are presumed to have a position of honor, perhaps in some relatively well-ordered and stable hierarchy of honor, though this assumption may be to a greater or lesser degree defeasable in light of the actual behavior of the agent. The notable in classical Greece having a public building constructed primarily out of a sense of abundance and power, if he is quite secure in his position, is unlikely to be focused upon the fact that his prestige is at risk of diminution, at least not until the psychological moment when he perceives things may be showing signs of going wrong.

What I want to suggest is that for a typical actor in the relevant culture, willpower, like honor among the Greek notables, is presumed on the part of the relevant actors. In the case of willpower, however, the status is distributed in a more egalitarian fashion. On the folk theory “every body has it,” so there is no reason to assume that one does not have it unless one has a proven track record of failure.

In the latter case, efforts of will take on the raw edge of naked self-assertion mingled with an air of desperation. At this point, one is more likely to be operating with the idea that “I will show them … .” Certainly such a person is regarded as “willful,”

---

though this term carries the connotation that something has gone wrong. I want to suggest that what has gone recognizably wrong is not just that there is an overabundance of willpower, but that the kind of over-zealousness in protecting one’s status against perceived slights that according to Benedict mainland Asian cultures tended to find alternatively barbaric and ludicrous, is also visibly at work here and is recognized for what it is. Unless such a person is young and trying to meet the challenges of passing into a position of responsibility, or has suffered some grave physical injury or calamity, or we for some other reason recognize his position as exceptional and feel some sympathy for his plight, we are likely to regard such performances as having a tinge of the pathetic and desperate. It is not that such an actor will not achieve his material purpose by such strategy, it is just that we are likely to see him as, to use more contemporary terms, “insecure,” perhaps profoundly so. Agents operating within the folk theory tend to have an idea of willpower as something more than this, and of being motivated by something other than what such transparent attempts to reclaim, demonstrate, or establish status or ability reveal themselves to be even to one operating naively within the folk theory. Except as one slides into doubt or desperation, the functioning of Status Staking may be obscured by the confidence of agents that anyone has the ability to exercise a vast or unlimited quantity of (metaphysical) willpower.

Consider again the diva. Even if her conscious motivation is to express something of beauty, and she does not focus on the fact that a certain amount of Status Staking is in play in her performance, nonetheless a desire to avoid a diminution of her status is an operative motive in her performance even if only as a kind of “backstop” if her energies are otherwise flagging. The Status Staking behind Output Willpower, I want
to argue, typically operates like the kind of Status Staking that occurs for the diva when she performs with confidence and without stage fright, it may be motivating, but can operate behind the scenes, so to speak.

More generally, an agent’s intent is to carry out some act, and he has some specific goal in mind in so doing. However, under the appropriate circumstances, such as when there are considerations which he knows suggest the task may be difficult to carry out because it is taxing, uncomfortable or frightening, by resolving to act he automatically stakes a status based upon a characteristic or ability that he is culturally imputed to have—willpower. This staking goes on automatically and, when we are operating from a position of relative confidence, neither the staking nor the stake need be conscious.

In support of this argument, first I note that performatives and the creation, maintenance and modification of institutional facts can occur without explicit knowledge or understanding of the process of the participating agents as is shown by the foundational work on this topic by Searle.\(^52\) This also seems to be the conclusion of anthropologists analyzing other cultures in terms of such concepts and processes.\(^53\)

Further, recent work in the psychology of action makes clear the great extent to which both motives and skills can operate unconsciously in remarkably complex ways.

\[^W\]hat the auto-motive model adds to the already extant and well-accepted notion of autonomous, well-practiced skills or goals is that the initiating act of will itself can become delegated to the environment. Take ... the example of driving ... . We have argued above that driving is a complex perceptual-motor skill, in which decisions as to how to move the wheel, how hard to push the accelerator, when to be ready to hit the


\(^{53}\) See e.g. Rappaport, Ritual and Religion.
brakes, and so on are guided non-consciously (in the experienced driver) by environmental information. In other words, these behavioral decisions are activated by information in the environment relevant to those decision processes. Now recall that those decisions, in the novice and less experienced driver, are made consciously. Therefore, with experience, decisions that used to have to be made consciously no longer are, and what makes those decisions if conscious processes do not? Those decisions as to what to do next—what subgoal to follow, in other words—are made directly on the basis of environmental information present. The information itself triggers those goal-directed actions.

Thus, in principle, there is no reason to believe that the goal “to drive,” or, to take a more social example, “to be patient,” cannot be removed from conscious control and delegated to the environment. This is the key hypothesis of the auto-motive model of unconscious motivations—that conscious intent or will can be bypassed, that the gap between the environment and the autonomous goal can be bridged, making the entire process from start to finish unconscious.  

If this is correct, then we might just as well consider a consciously controlled event, such as a deliberately undertaken action, to be an environmental cue no different from any other vis-à-vis the unconscious goal-directed process. Thus, staking as a goal-directed process could operate in precisely this way.

As for when the “skill” or mechanism is consciously practiced such that it might become automatic, it seems that the story we want to tell is that, like most such institutional practices and facts, we are inculturated into them from early on. It seems plausible that the establishment of the mechanisms of praise and blame that are connected with attributions of willpower and that ultimately become internalized would be inculturated at a quite early age, probably before a great deal of reflective awareness and an ability to make fine discriminations in motivations is generally operative. Of course, making this developmental account anything more than a plausible speculation would

---

require a great deal of empirical work, and, again, I must beg off and rely on others for their efforts and results in this regard if this idea is ever to be put to the test.

Once a metaphysical theory of willpower is placed on top of the actual experiences of the agent, the actual operations of the underlying mechanism may not only be outside of focal awareness, but awareness of them may actually be suppressed. Appeals to a mechanism may be perceived as tantamount to denial of responsibility for one's action, and as such may perceived by the agent or others as threats to the evaluative system. Tales of how this happens within cultures, institutions and families are commonplace in the anthropological and psychological literature, and I do not intend to defend a particular theory on this score, but merely advert to the general acceptance of the fact that such processes occur. Crucially, it should be noted that on any explanation of Output Willpower consistent with the BMA, the agent operating on the folk theory to the extent that he ascribes to a metaphysical theory of will *must* be mistaken, and, in order for this mistake to be durable, not conscious of the real motives and mechanisms driving the behavior that he associates with Output Willpower. For no motive or mechanism acceptable under the BMA could answer to the description of the kind of causally potent motivations arising *ex nihilo* that are ascribed to the exercise of Output Willpower under the folk theory.

Assuming the motivations of Output Willpower are unconscious, post hoc explanations of action may then be constructed by agents in the only terms readily available to them—the terms of the metaphysical theory of willpower. The power that such post hoc rationalizations may exert is strikingly displayed in the following passage:

Gazzinga has noted the phenomenon [of the creation of *post hoc* explanations for action] in split-brain or Korsakoff's syndrome patients: A

177
message is flashed on their right brain hemispheres to get up and leave the room (for example), and they do so. When stopped by the experimenter and asked where they are going, the subject responds nearly immediately with a plausible conscious motive, such as “I needed to get a drink of water.” And posthypnotic suggestions have the same flavor. A subject is given the command that when she awakens from the trance, she is to crawl around on the floor on her hands and knees. She is awakened; she crawls around on the floor and says, “I think I lost an earring down here.”

In exercising the Status Staking mechanism subtending Output Willpower, what the agent may experience in focal awareness is simply that he makes a commitment to action followed by—presto—a boost in motivation to achieve whatever he had set his mind to do. It seems to him like magic, or, if you like, a very special power different from other motives on the basis of which he operates.

I think this phenomenological point can be given some substantiation by comparing the seemingly magical causal powers attributed both to the properties of the will and to the properties of the performatives under folk theory, and the associated feelings which can retain at least some of their vitality even when we have some understanding of the underlying mechanism. O’Shaughnessy writes in his monograph on the will:

Then first thoughts concerning the will tend to conceive of it as a kind of magical emission from the psyche, an occult mental electricity or magnetism that originates in an event of choice that is directed onto entities situated at a remove from the mind, manifesting its occult power in the most striking and doubtless the original case of ‘action at a distance’. For if the magical situation is one in which nature at large is immediately responsive to the mind, so that the full autonomy and distinctness of physical objects generally, and so also of the self, is improperly established, and the ‘sense of reality’ and rationality of the practitioners is not total; if it is that state of affairs in which objects of their nature inexplicably respond to mental events, rather as if distant objects were not wholly separate from one and could listen in to and hear and even obey one’s very thoughts!; then I suggest that in converse

---

55 Ibid. at 463 (cites omitted).
manner the relation that obtains in willing can at first blush appear magical. Instead of saying that the magical is to be construed as an extension of willing, one might wish rather to say that willing should be understood as a localization of the magical. So, at any rate, it seems to me. Wittgenstein remarked that we would liken the mind to a gas-filled bowl did we not feel so foolish in doing so. Then it seems to me no less natural to think of the relation between some decision event and toe movement, not merely as a thing to marvel at, but as somehow breaking all the rules! beyond the pale of reason! in a word, akin to magic! A grossly false idea no doubt, primitive even; but, I submit, decidedly natural. ‘Philosophical repression’ may attempt to laugh it off, but such defensive thought tends to generate its own retributive punishment (what Freud called ‘the return of the repressed’). In any case, I shall take seriously this comparison between the will and a magical force.56

This kind of experience of the peculiar force involved in willing in general, I have tried to show, is involved in Output Willpower in rather specific ways.

Rappaport notes the connection between the sense of an occult power or force and performatives in explaining the occult efficacy of ritual performances. First, Rappaport offers an explanation of the “occult” as follows:

If ritual (in contrast technique) does anything at all it doesn’t do it by operating with matter and energy on matter and energy in accordance with physical laws of cause and effect, but by focusing agencies or forces of another sort upon whatever is to be affected.

This difference is likely to be understood by the actors who, although they may not make a clear distinction between what is natural and what is not, know the efficacy of spells to be different from that of spears. Terms like “supernatural” and “preternatural” have been used to refer to the non-physical agencies posited by people and invoked by them in their rituals. But Fortes, for one, objected that the terms “supernatural” is an artifact of literate cultures, and claimed that the actor, in tribal societies at least, sees the world as composed of the patent and the hidden—or occult—which present themselves in mixed sequences and are to be interwoven into a unified reality. He took this distinction to be widespread, or even universal. The patent can be known in the last resort by sensory experiences and it confirms to the regularities of material cause. The occult, presumably, cannot be so known nor does it so conform. Ritual, in

---

his view, is distinguished from non-ritual, not only by the analyst but by
the actor as well, in that it derives its efficacy from the occult, and not the
patent. 57

Rappaport later describes the connection between performatives and occult powers.

[A]s Ruth Finnegan has suggested, albeit rather nonspecifically, the “truth
lying behind” assumptions concerning what is often called “the magical
power of words” may be related to their illocutionary force or
performativeness. It may be proposed, rather more specifically, that the
magical power of some of the words and acts forming parts of liturgies
derives from the relationships between them and the conventional states of
affairs with which they are concerned. As we already observed, the
relationship of performatives to the states of affairs with which they are
concerned is the inverse of that of statements. The facts, events or
situations to which a statement refers presumably exist independent of and
previous to the statement referring to them, and a statement is assessed
ture if it accords in some sufficient degree to those previously existing
independent states of affairs. Since performatives bring about the facts,
events, situations, etc. with which they are concerned, these facts are
subsequent to and contingent upon them. Performatives, and most
unambiguously factives, are self-fulfilling: they make themselves true in
the sense of standing in a relationship of conformity to the states of affairs
with which they are concerned. In light of this it may be suggested that
the performativeness, and especially the factiveness of ritual acts and
utterances provides a basis for occult efficacy in general, including the
magic powers of words in particular. Ritual’s words do, after all, bring
conventional states of affairs, or “institutional facts” into being, and
having been brought into being they are as real as “brute facts.” It may
also be that magical power is attributed to other words by extension of the
of the principle of factiveness beyond the domain of the meaningful, in
which it is clearly effective, into the physical, in which it is not … 58

Thus, there are some grounds for seeing magical force associated with the will as
a special instance of the attribution of occult powers generally connected with the naïve
use of performatives.

In fairness, it may be that agents are neither totally naïve to the operation of
willpower nor completely within the sway of the hypothesis of an underlying causally

57 Rappaport, Ritual and Religion, 48 (cites omitted).

58 Ibid. at 117.
normal mechanism. It may be that the agent operates in a kind of psychological half-way house between genuinely and fully believing in a metaphysical willpower to the exclusion of all other explanations, and something less demanding. Bratman, for instance, has argued that acceptance, as a psychological state, must be clearly distinguished from belief, and that it may be typical and rational to accept certain things as correct or true in various social or practical contexts, even though they do not command our belief.⁵⁹ Taking another tack, other authors have argued that things can be believed in quite different ways as appropriate to their objects, religious as opposed to practical matters for instance, so that apparently incompatible beliefs co-exist peacefully in one agent.⁶⁰ Complicating matters further, and creating yet more psychological space for officially disclaimed concepts (such as metaphysical willpower) to hide, is the fact that we may often be somewhat confused about the fact of the matter of what we actually believe.⁶¹ Hence, there are a variety of possibilities for the agent’s attitude toward metaphysical willpower, while at the same time, with more or less naiveté about the underlying mechanism, he may experience the process of exercising willpower as in important respects involving unique motivations peculiarly responsive to his own choice to engage it.

---


IX. **Summation**

If my arguments are correct, I have presented an explanation of Output Willpower that meets the criteria set out in Chapter Two. Specifically, I have set out an account of a mechanism that enables us to (1) overcome temptation (2) while maintaining a clear focus on just what the primary rewards at stake are and are having no real awareness of any added stake, except in marginal cases, (3) despite the fact that there is an added motivation (4) that is apparently *sui generis* in being responsive only to one's determined decisions to act and that is not traceable to or connected with any discernable external consideration, and is the same motivation in all relevant cases. As I say, this motivation results from (5) one's decision to act in a determined manner in spite of temptations and obstacles, providing a wholly "internal" means of resisting temptation that (6) is a causal force operating in behavior (7) that can be engaged as a matter of degree as the circumstances or the amount of status staked dictates, but which (9) is engaged in a manner unconnected with ethical or rational mandates as these are conventionally understood. By the same token, in betting a single, one-time stake, neither the One-Shot Objection nor the Unraveling Problem has any purchase on this mechanism, and I have explained at length how the relevant staking is self-enforcing, so the Bootstrapping Problem does not threaten.

X. **General Objections, Conclusions and Prospects**

I do not consider the account that I have offered to be necessarily an exclusive account of Output Willpower. Initially, I had hoped that this account to be the single, unique explanation of Output Willpower could be proven. However, the further I went in examining and ruling out alternatives, and in examining the rich empirical literature
relevant to the topic, the clearer it became that I had little grounds for ruling out in
dvance potentially unforeseen competitors through mere conceptual argument. In the
end I realized that such efforts were, in fact, against the whole spirit of the project.
Perhaps equally important, I gained a sense of perspective about Output Willpower—
realizing that it is only one of a great many interesting techniques that may be employed
to overcome temptation or enhance motivation and that may or may not travel under the
name “willpower” in everyday language.

This is not to say that I take the effort to understand the mechanism of Output
Willpower to be unimportant. Having surveyed, assessed and discarded numerous
explanations and half-explanations of willpower constructed by other authors and by
myself, I have found that elucidating any decent explanation in this area is extremely
difficult, and the attempts thus far have been gross failures. I have tried to demonstrate
this with respect to the mechanisms surveyed in this dissertation. The only mechanisms
that even begin to provide a prospect for an alternative explanation of Output Willpower,
other than variations on the Personal Rules mechanisms, are the kind of reframing
described by Taylor,62 and some form of verbal or mental self-stimulation of the
motivational “nodes” associated with a given course of action. With respect to the
former, I have described in Chapter Three how such would be subject to the Device
Objection. Further, in fundamentally changing the terms of the choice through reframing
the decision, it seems as if the Clear-Eyedness Requirement would also be violated. With
respect to the latter, we are essentially talking about what amounts to “psyching oneself

up" through e.g. "inner speech." Again, the Device Objection immediately poses a problem. Further, for both of these alternatives the Ordinary Motive Objection and the Multiple Motives Objection would also be telling unless the underlying motivation stimulated or brought into focus by reframing was something like the motive I have posited as operative in Output Willpower conceived as Status Staking. After all, this appears to be the only candidate for a motivation that is operative in all and only those cases where one is exercising Output Willpower. But if this motive is assumed, then it is hard to see why this kind of self-stimulation or reframing would be attractive as explanations for anything other than a mechanisms ancillary to Status Staking.

There is one general objection to Status-Staking as an explanation of willpower that is well worth considering. That is: Why not suppose that Output Willpower can be explained simply in terms of a strong meta-preference to do whatever we have set out minds to doing? Why take the extra step and hypothesize the complicated structure of pledging? After all, such a meta-preference would apply itself as a motivational force in all and only those cases where we resolve to take action—just as Output Willpower does on the folk theory.

I think that the meta-preference view fails for several reasons. First, I think it implausible that such a meta-preference standing alone would be strong enough to provide the enormous motivational power often associated with acts motivated by Output Willpower. Second, the meta-preference view has no resources to explain the fact that Output Willpower has the potential to be engaged to varying degrees with respect to the same objective. In other words, the meta-preference view presupposes a single constant

---

meta-preference, which is presumably stable in its strength or motivational power, and it is not clear how such a fixed motivation could produce various degrees of effort. Third, the meta-preference view does not have the resources to explain the proportionality between the strength of the resolve to take action and the strength of the motivations operative in Output Willpower. We might hypothesize a tailor-made meta-preference to carry out the actions one resolves to undertake which is sensitive to the degree of firmness in one’s resolution, but this seems rather ad hoc, and it is not clear how exactly this would operate. For the strength of the motivation operative in Output Willpower is not proportionate to how much one otherwise wants to achieve the aim—so that the meta-preference could operate, so to speak, sympathetically, just as one naturally wants to help another to achieve his aims to a greater degree where it is clear the he strongly desires to achieve those goals. Rather, the added motivation operative in Output Willpower is geared to level of commitment signaled—perhaps signaled only to oneself— as opposed to felt or emotively expressed. It is not clear that one could tie the supposed variable strength of the meta-preference to such a signal without supposing that there is some side stake at issue that is pledged via the signal. Fourth, in the cases where failure of willpower threatens reveal the “raw edge” of the operative motivation, something else is at stake in the operation of will, something intimately connected with the agent’s sense of himself and his own power to effect his aims. That is what gives the agent’s action its sense of desperation—as has already been discussed. Such a stake cannot be explained merely by a meta-preference to carry out one’s committed aims. Fifth, the cost of an actual, as opposed to merely threatened, failure of willpower reveal

––––––––

64 See supra. Section VII.C.
that a predominant stake in the success of such Output Willpower-motivated action is the agent’s own sense of his power and effectiveness. The agent who has suffered a failure of Output Willpower is likely to suffer a loss of faith in himself and his ability to control his own actions, and a concomitant sense of failure about himself as a powerful and effective agent. This is more than frustration at not having one’s meta-preference satisfied. These emotions speak to what kind of agent one takes oneself to be, and this, I suggest, is the dominant stake in such a decision. This stake is more fully revealed when it figures as a cost of failure of Output Willpower. A meta-preference alone cannot explain this cost of failure.

This is not to say that no meta-preference is operative in exercise of Output Willpower. Indeed, it seems plausible that a free-standing meta-preference to live up to one’s commitments to action would be operative in agents whose status is staked upon living up to such commitments. In some cases, this meta-preference alone may be sufficient to provide the requisite motivation for success of action. Further, to the extent that one is conscious of such a motivation, the presence of the motivations associated with Status Staking can more easily be kept in the psychological background, so to speak, outside of conscious awareness. However, in order to plausibly explain Output Willpower, another source of motivation, such as that provided by Status Staking is clearly required.

I hope that I have made at least some progress in explaining Output Willpower, and perhaps have even substantially explained the target phenomenon. Of course, the final demonstration of whether Status Staking explains Output Willpower will depend in substantial measure upon empirical investigations which I am not equipped to conduct. I have merely tried to demonstrate that Status Staking is a possible explanation in that it
could operate in a manner that is functionally isomorphic with the purported operations of Output Willpower, and that it could be compellingly perceived by agents as the operation of Output Willpower. I have also gone to some lengths to show that in general terms it is a psychologically plausible theory. I am aware of no competing theory that comes close to matching these explanatory virtues.65

XI. Conclusion

In explaining willpower as Status Staking, I think I have given it its fair due. By virtue of explaining Output Willpower as a form of Status Staking, we are able to avoid simply sweeping it under the carpet as a theoretical embarrassment or mere figment of folk imagination. Rather, I have shown it to be a vital mechanism, causally effective in motivating our behavior. In so doing, I hope to have given form and substance to the reality lying behind experiences and intuitions informing the folk theory. I think that this is the best we can do for the folk theory. To go any further would be to accept a magical account of volition that would be an anathema to any scientific account of behavior, and relegate this important phenomenon, so vital to the understanding of human behavior in all its possibilities and pathologies and so central to our concept of autonomy, to the realm of mere myth. Science and philosophy have for a long time attempted to ignore willpower or define it out of existence, but it has always defiantly remained with us, refusing to fade away into a mere quaint, though irrelevant, fiction.

---

65 A topic worthy of mention, but one that I will have to set aside, is what to make of pathologies of the will. Given the remarks I make in Chapter One, this is obviously a topic that interests me, just as it interests James and Ainslie. William James, The Principles of Psychology, 2 vols., (New York: Henry Holt & Co. 1890, reprint Dover Publications, 1950), Vol. II., 522-569; Ainslie, Picoeconomics, at 154. However, I must save that investigation for another day.
Bibliography


Alcoholics Anonymous World Services, Inc., *Alcoholics Anonymous*, 1986


Hardin, Russell, Collective Action, (Baltimore: Johns Hopkins University Press, 1982).


196


Pears, David, Motivated Irrationality, (South Bend, Ind.: St. Augustine’s Press, 1998).


Appendix A

The Iterated Prisoner's Dilemma Analysis of Personal Rules

Ainslie argues that Personal Rules can be understood in terms of an iterated prisoner's dilemma analysis. While Ainslie does not specifically link this analysis to *Prescriptive* Personal Rules, in a somewhat confusing mix of metaphors, he does connect the idea of an iterated prisoner's dilemma to "promises" which are "self-enforcing."¹ Therefore, Ainslie may take the analogy to prisoner's dilemma to support his view that Prescriptive Personal Rules are similarly self-enforcing. Before assessing Ainslie's analysis, it is best to first review the concept of a prisoner's dilemma. I will first explain what a prisoner's dilemma is. I will then set out the reasons why game theorists believe that cooperation occurs in an iterated prisoner's dilemma. Having set out the argument regarding the iterated prisoner's dilemma, I will then draw out Ainslie's analogy in more detail, with specific attention to the nature of the "players" in the iterated prisoner's dilemma. In so doing, I will explain why Ainslie's analogy must fail.

The prisoner's dilemma is represented by a two-by-two matrix depicting four possible outcomes where each of two agents, or players, has to choose between two alternatives. (See figure A.1²). The interest in the prisoner's dilemma arises from the fact that, given the nature of the choices presented, if each player chooses that which is best for himself, they both end up worse off.³ This can be made clearer by giving a concrete


### Figure A.1

<table>
<thead>
<tr>
<th>You</th>
<th>do (1)</th>
<th>do (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>do (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>do (2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Third-best for each</th>
<th>Best for me, worst for you</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worst for me, best for you</td>
<td>Second-best for both</td>
</tr>
</tbody>
</table>

### Figure A.2

<table>
<thead>
<tr>
<th>You</th>
<th>confess</th>
<th>keep silent</th>
</tr>
</thead>
<tbody>
<tr>
<td>confess</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>keep silent</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Each gets 10 years</th>
<th>1 go free, you get 12 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>I get 12 years, you go free</td>
<td>Each gets 2 years</td>
</tr>
</tbody>
</table>

202(a)
example—the example for which the prisoner’s dilemma was named. Parfit captures the matter most succinctly.

You and I are questioned separately about some joint crime. . . .

Whatever the other does, it will be better for each if he confesses. By confessing each will be certain to save himself two years in prison. But if both confess that will be worse for each than if both keep silent.  

The payoff structure is depicted in figure A.1.  

Having seen this specific example, it should be easier to grasp the general form of the payoff structure for prisoner’s dilemma as depicted in figure A.1. Whenever this payoff structure is present, the following ”dilemma” arises: if each player chooses that which is best for himself, they both end up worse off. If we assume that agents will tend to choose that which is best for them, then situations that may be formally represented as prisoner’s dilemma payoff structures will tend to yield these tragic results.

A key feature of the prisoner’s dilemma is that one agent’s choice does not affect that of the other. For if Player One were able to punish or reward Player Two for confessing (“defection”) or staying silent (“cooperation”), then it might no longer be in the interest of Player Two simply to defect without regard to Player One’s behavior. In other words, defection would cease to be the “dominant strategy.”

So far, I have only discussed single-play prisoner’s dilemma. However, players may face each other repeatedly in situations with prisoner’s dilemma-type payoff structures.

---


5 Adopted from Ibid. at 57.


When the same players encounter one another repeatedly in prisoner’s dilemma scenarios a change in incentives occurs so that defection is no longer the dominant strategy. The solution to the problem of mutual defection in an iterated prisoner’s dilemma takes the form of tacit cooperation. This is because some form of tacit cooperation could be established between the players in the prisoner’s dilemma through the following strategy. A player can punish defection on the part of the other player with defection on his own part, and reward cooperation on the part of the other player with his own cooperation. By so doing, a player can create the expectation in the other that defection will be answered with defection and cooperation with cooperation. Obviously, this strategy is open to both players. Through this sort of tacit coordination, cooperation may emerge in the iterated prisoner’s dilemma. This argument is well known amongst game theorists.  

Ainslie likens the self-enforcing quality of a Personal Rule to an iterated prisoner’s dilemma. This analogy is hopelessly flawed. In attacking this analogy I will not challenge the game theorists’ argument that in iterated prisoner’s dilemma cooperation will naturally tend to evolve. Rather, I will argue that Ainslie’s analogy to iterated prisoner’s dilemma fails because Ainslie fails to tell a coherent story about who the “players” are who repeatedly face each other in the prisoner’s dilemma, and even when we supplement his account to provide such a story, it doesn’t really explain Prescriptive Personal Rules. Therefore, he can make no use of the game theorists’ argument to prove that Prescriptive Personal Rules are self-enforcing.

A key problem in unpacking Ainslie’s analogy to the iterated prisoner’s dilemma is that he never tells us who the players in the prisoner’s dilemma are, other than that they are

---

<table>
<thead>
<tr>
<th>Now</th>
<th>Future</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stay Up</td>
<td></td>
<td>To Bed</td>
</tr>
<tr>
<td>Stay Up</td>
<td></td>
<td>64 + 14 = 78</td>
<td>64 + 56 = 120</td>
</tr>
<tr>
<td>To Bed</td>
<td></td>
<td>49 + 14 = 63</td>
<td>49 + 56 = 105</td>
</tr>
</tbody>
</table>

Table A.1 Present Value of Present and Future Choices
generally “successive motivational states.” Nonetheless, we can get a start on unpacking Ainslie’s analogy by looking at the prisoner’s dilemma matrix he uses to illustrate his theory. Ainslie illustrates his intrapersonal prisoner’s dilemma in Table A.1.⁹

Before describing who the players in this prisoner’s dilemma might be, I will first describe why Ainslie sets the values he does on the choices illustrated by the matrix. The matrix is intended to illustrate the case of the individual discussed in Chapter Three who nightly faces the choice of either staying up an extra two hours to watch TV or going to bed early in order to be well-rested for the next day’s work. All choices are calculated on the basis of their present value, at midnight, when the individual actually makes the first choice in the series.¹¹ 64 represents the present value of staying up late tonight.¹² 49 represents the present value to the agent of going to bed early tonight and being well rested tomorrow.¹³ Thus, if the choice was made for only one night, the agent would choose to stay up and watch television for an extra two hours.¹⁴ However, it is also the case that the present value to the agent of going to bed early on the nine nights after this one is 56, and the present value for staying up is 14.¹⁵

---


¹¹ I will not describe further Ainslie’s calculations for deriving these values as it is irrelevant to the argument. Ainslie’s argument fails regardless of how he arrives at his assignments of value.

¹² Ibid. at 191.

¹³ Ibid. at 191.

¹⁴ Ibid. at 191.

¹⁵ Ibid. at 193.
The rationale for assigning these values is Ainslie’s theory of the hyperbolic
discounting of reward value over time.\textsuperscript{16} The reason staying awake tonight has a high
reward value is that the reward for staying awake occurs in the immediate future, while the
enjoyment of being well-rested tomorrow is significantly discounted because it will not
occur until the next morning. However, as one looks further into the future, the shape of the
discount curve gives a higher present value to the future incidences (beyond this evening) of
going to work rested than it does to future incidences of watching television. Regardless of
the correctness or incorrectness of this theory of discounting, Ainslie must explain how it is
that an agent can choose to create the expectation that these rewards will stand or fall
together—ten nights of staying awake late or ten nights of going to bed early—by means of
a Personal Rule and how it is that the analogy to iterated prisoner’s dilemma is supposed to
help.

Since Ainslie never clearly explains who the two players in Table A.1 are, we must
do a bit of “reverse engineering” to understand what he intends to depict by the payoff
structure in this matrix. The first question is whether the matrix is intended to depict the
payoff structure for a single player—i.e. a single “motivational state”—or the payoff
structure for two such players. The matrix cannot represent the payoff structure for two
players, i.e. with the first addend representing the payoff for Player One and the second
addend representing the payoff for Player Two with their sum equaling the collective payoff
for the “whole” individual who has the conflicting “motivational states.” This is because
interpreted this way the matrix simply does not represent a prisoner’s dilemma. That is, on
the two-player interpretation, no matter what Player Two does, Player One should choose to

\textsuperscript{16} \textit{See infra} Chapter Three, Section II.
stay awake tonight, and no matter what Player One does, Player Two should choose to go to bed on the next nine nights. Thus, each chooses the optimal outcome for him and each enjoys the optimal outcome. Neither causes the other to suffer by choosing what is in his own best interest. The dominant solution is the best outcome for both players.

The only other possible interpretation of the payoff structure is that the sum of the two addends represents the respective payoffs in the payoff structure. In that case, the matrix shows the payoff structure for a single player, call him “Player One.” Viewed in this way, Table A.1 does represent a payoff structure for a prisoner’s dilemma. Consider staying awake to be “defecting” and going to sleep to be “cooperating.” Clearly, no matter what any other player does, Player One should defect. If the next player, Player Two, faces a similar payoff structure, then he too will defect. Thus, night after night the agent, comprised of two “successive motivational states,” will succumb and give into temptation. Thus, at each stage, the agent would choose to defect now and cooperate later, but when “later” comes, the agent will prefer to defect. In this sense, the agent does repeatedly act in an iterated intrapersonal prisoner’s dilemma.

It is also true that, if constrained to choose between ten cooperating choices and ten defecting choices, the agent would choose the ten cooperating choices—105 is better than 78. However, the mere fact that it would be better according to the agent’s current reward function for him to choose ten cooperating decisions over ten defecting decisions does not mean that it is in his power to determine now the outcome of those later choices. This is the Bootstrapping Problem all over again. What is needed is something to bind those decisions together. The fact that it would be beneficial if they were so bound together is not enough. Therefore, Ainslie needs to give an account, not of a single player who repeatedly plays in
prisoner's dilemmas, but an account of an iterated prisoner's dilemma—i.e. a series of plays where the same two (or more) players play each other repeatedly so that cooperation can evolve.

We then return to the question: who are Player 1 and Player 2? First, Ainslie tells us that they are "successive motivational states." However, Ainslie does not clarify just what he means by a "motivational state." Presumably he means a given set of desire tokens. Unfortunately, with this definition, the analogy to iterated prisoner's dilemma cannot work. Consider the person deciding whether to stay up late and watch TV. After tonight, the desire to stay up on this particular night will be gone—whether the agent stays up or not. The desire will either be satisfied or will naturally disappear since it will no longer have an available object. Another desire of the same or similar type may occur on the next night, and many nights thereafter, but they will be different desire tokens. The same is true with the desire to go to work well rested in the morning. Therefore, none of the agent's successive motivational states will be the same motivational state that exists tonight. Of course, the same is true with respect to each succeeding motivational state and its respective successors. Therefore, there are no repeat players between whom cooperation can evolve.

To get around this problem, Ainslie might argue on the basis of a different class of cases, namely, that wherein one or more of the desires comprising the successive motivation states persist through several such successive states. Thus, for instance, and alcoholic may have numerous successive desire tokens which are desires to drink, but may have a continuous desire to enjoy the benefits of a sober life, or to be, or to be considered to be, a

sober person. But this does not help. The motivational state at t1 is still different from the motivational state at t2 because one or more different desire tokens are operative even though the general desire for sobriety remains constant throughout. For a new desire token for taking a drink will have formed and the old one will have passed away. Hence, there can be no reasonable grounds for saying that the configuration of motivations at t3 (with or without the continued desire for sobriety) will be the same motivational state as existed at t1. So, again, it is not a case where two players face each other repeatedly.

Conceivably, Ainslie could argue that there might be an ongoing battle between two such persisting long-term desires—perhaps a desire for a cautious, sober life and a desire for a carefree, hedonistic life. In this case, successive motivational states could be individuated on the basis of which of these long-term desire tokens was dominant. There is some textual support for this interpretation in Ainslie’s analogies to alternating majority parties in a legislature. However, none of Ainslie’s concrete examples of willpower or of failures of will fit this paradigm. This is probably because this model does not fit the usual cases of failure of will very well, where temptations are typically driven by short-term, transient motivations. Nonetheless, even if we assume that such recurrent preference patterns could emerge, any “cooperation” that emerged between them would have nothing to do with Prescriptive Personal Rules. Cooperation would emerge tacitly through repeated iterations, and there is neither a need nor a place for Prescriptive Personal Rules in such an explanation. Cooperation emerges on the basis of the facts as they stand.

---

18 Ibid. at 189.

19 Ibid. at 191.
A last possible defense of the analogy to prisoner's dilemma is that it is "present self" vs. "future self" (or present self vs. present self's imagined future self). Indeed, Ainslie's labeling of Table A.1 lends some support to this view. But, again, this does not help matters. "Present self" unless it is a mere abstract category, is merely another name for the occurrent set of desire tokens at a time t1, and faces precisely the same line of objections as was set forth above. If "present self" simply designates whatever mental and motivational states are occurrent for the agent in the present moment, so that it functions as a kind of indexical—the "now-self" so to speak, then "present self" exists over time only as an abstract designation, and qua abstract designation it has no interests, and cannot be a player in the prisoner's dilemma. Therefore, on any available interpretation, the analogy to iterated prisoner's dilemma fails to support Ainslie's hypothesis of the existence of Prescriptive Personal Rules and their ability to explain Output Willpower.

If we are to make any sense of Personal Rules in terms of prisoner's dilemma we must not look to the model of an n-person iterated prisoner's dilemma as suggested by Ainslie, but rather to a very different model found in the work of Elster.

Elster makes the following observation:

The paradox of democracy can be thus expressed: each generation wants to be free to bind its successors, while not being bound by its predecessors... [I]t is possible for any given generation... to eat its cake and have it, but all generations... cannot simultaneously achieve this goal. The link to the problem of time preferences... is obvious. 20

Unfortunately, Elster apparently feels the link to the problem of time preferences—\emph{i.e.} to what I have called "time discounting"—is simply too obvious to set out and so he does not do so. This task remains to be carried out.

The idea here is that each generation ideally would like to establish its political program based upon its own preferences and see this program carried on into the future. At the same time, that same generation will not want to have the effectuation of its aims hampered by legal or other restraints placed upon it by past generations. One element that powers this "paradox" is that a political actor often wishes enact political measures that will have effects upon later generations, or the world at large, even after the point at which he believes he will no longer be in existence to enjoy their fruits or see them carried out. This may be an expression of values or ideals, or simply a desire to have things turn out in a certain way. But this is a political commonplace.

Similarly, a person may, at a given time, have a desire that he wishes to have fulfilled in the future, regardless of whether he (or the current "person-stage" or current "motivational state") is in existence at the point in time when the event which would fulfill the desire transpires. Parfit describes such desires as an instance of the class of desires that are not conditioned on their own persistence.\footnote{21 Derek Parfit, \textit{Reasons and Persons}, (Oxford: Clarendon Press, 1984), 151.} He offers the following discussion:

\begin{quote}
... Suppose that I meet some stranger on a train. She describes her life’s ambitions, and the hopes and fears with which she views her chances of success. By the end of the journey, my sympathy is aroused, and I strongly want this stranger to succeed. I have this strong desire even though I know that we shall never meet again. This desire is not implicitly conditioned on its own persistence. The same is true of other desires of many kinds.
\end{quote}
The clearest examples are the desires people have about what will happen after they are dead. Suppose that I do not believe that I shall have an afterlife. Since I believe that death will be my extinction, my desires about what happens afterward cannot be conditional on their own persistence up until the time of their moment of fulfillment. I believe that this condition cannot be fulfilled; yet I still have these desires.\(^{22}\)

The point here is simply that there are desires the existence of which does not depend upon the agent, person-stage, motivational state, or even the instant desire itself, being in existence at the point in time when the event which would satisfy the desire will occur.\(^{23}\)

Now the analogy to the kind of case Ainslie is worried about—repeated exposure to a given class of temptations—should become clearer. Each “successive motivational state” would seek to have its own preferences, even those not conditional upon their own persistence, fulfilled and so, implicitly or explicitly, bind future motivational states. At the same time, each such “motivational state” or “person-stage” does not wish to be bound by the preferences of the past person-stages in effecting its aims. Tonight I want to stay up and watch TV late this evening and to go to bed early on future nights—I do not care that last night I wanted it to be the case that I would go to bed early on the evening that has now arrived. Thus, the prisoner’s dilemma structure here is not iterated in the ordinary sense. Rather, it is as if Player One makes decisions affecting Player Two, Player Two makes decisions affecting Player Three, and so on.

Player One and Player Two never exist at the same time. Therefore, strictly speaking, Player Two can neither assure nor threaten Player One with respect to Player

\(^{22}\) *Id.*

\(^{23}\) Whether these desires ought to be given weight in a normative theory of rationality is a separate issue that Parfit goes on to consider at length. See e.g. *Ibid.* at Sections 59, 61. However, this issue needn’t trouble us here, as I only need to establish that such desires exist, not whether and how they should be assessed in a normative theory of rationality.
Two’s moves in future rounds, for Player One will have passed out of existence before Player Two has the opportunity to do so. Hence, no tacit cooperation can develop between them. However, Player One does have something to lose in that Player Two can prevent the fulfillment of those of Player One’s desires that are not conditioned on his (or their) persistence. Therefore, if Player One believes that his own cooperation or defection will increase or decrease the likelihood of Player Two’s cooperation, then Player One has a reason to cooperate. The question is: How could Player One’s decision to cooperate or defect influence Player Two’s decision? Ainslie gives us no clear answer. For one thing is certain, it cannot be through Player One’s assurances of future cooperation or threats of future defection as is the case with iterated prisoner’s dilemma, for Player One will not be around in the future to offer rewards or to mete out punishments.

Still, one possible route lies in Ainslie’s idea of a behavioral “precedent.” He writes that “… players in experimental bargaining games, such as prisoner’s dilemma, are motivated to choose the cooperative solution if that choice will be seen as a precedent for future games, but will often choose the uncooperative solution if it will not.”24 The notion of precedent is present in a number of places in Ainslie’s work. I would suggest that, as with so much in Ainslie’s work, he is incautiously mixing his metaphors or analogies without carefully working them out—but the seeds of interesting and important ideas lie there nonetheless.

If we take the idea of precedent seriously, then, at least in the legal setting, it means a decision that establishes, changes, develops or modifies the existing standard as

---

to what is correct action in a given instance. A decision that it is reasonable or permissible to fail to perform a contract under circumstances C1 implies that it is permissible in the future to fail to perform a contract under circumstances relevantly similar to C1. In the everyday life of an agent, a decision to carry out one course of action rather than another may tend to legitimize that course of action—e.g. “this is the way I have always done it, and that is good enough for me.”

Alternately, choosing a given course of action, for instance, may tend to create, further or strengthen a given self-conception, and this self-conception may in turn tend to make the agent more likely to make similar decisions in the future so as to validate this same self-conception. For example, if I have told the truth under difficult circumstances, this might further my self-image as an honest man, which in turn may create an increased stake in my acting this way in the future in order to preserve or enhance a self-image in which I have become increasingly more invested as it grows in strength or scope as a result of my prior acts. Thus, there are a number of causal mechanisms that may link my choice in the current instance with choices in relevantly similar future instances such that my choice in the current instance will tend to promote a similar choice in those future instances, and such that the recognition of this fact will provide an increased motivation in the current choice.

However, the ultimate point is that some kind of causal or otherwise fixed connection must exist between the decision points in order for them to hang together. Rightly understood, the prisoner’s dilemma analogy does nothing to explain what this link is vis-à-vis Personal Rules, let alone show how it could be established by fiat as is
supposed by the theory of Prescriptive Personal Rules. The prisoner's dilemma analysis does nothing to support the hypothesis of Prescriptive Personal Rules.