WILL AS INTERTEMPORAL BARGAINING: IMPLICATIONS FOR RATIONALITY

GEORGE AINSLIE[†] & JOHN MONTEROSSO[††]

Rationality has been understood as conducting yourself according to reason rather than passion. In modern times this endeavor has become synonymous with maximizing your expected goods, with the value of expected but delayed goods discounted exponentially. However, behavioral research has found a robust tendency for delayed goods to be discounted hyperbolically, that is, for their value to be divided by their delay. This finding supplies a simple hypothesis about the origin of irrationality, but greatly complicates the problem of rationality, since it depicts a limited warfare relationship among interests within an individual. Recent research on the combining properties of hyperbolically discounted rewards supports the hypothesis that a person's will arises from a prisoner's-dilemma-like relationship among successive motivational states. This picoeconomic hypothesis provides a mechanism for both the strength and "freedom" of the will, and predicts pathologies of overcontrol that make strength of will something very different from pure rationality. This approach offers insights into current puzzles about criminal responsibility and the disease model of addiction.

INTRODUCTION

Rationality is an ancient concept, one that Plato contrasted with passion to form a dichotomy of choice principles.[1] Through the ages rationality has meant the good way to make choices, the way that will maximize your satisfaction with the outcome. As such, it has been a norm rather than a description of actual behavior. However, since utility theory has postulated that people always maximize their expected utility, rationality has acquired a descriptive implication: the rational is what anyone inevitably will do whenever she is aware of the true contingencies she faces. Irrationality then is merely error, the product of some fallacious valuation process. Modern rational choice theory (RCT) thus aims not only at normative optimality but also at factual accuracy; it consists of "a series of assumptions about

---

[†] Chief Psychiatrist, Veterans Affairs Medical Center, Coatesville, Pennsylvania, and Clinical Professor of Psychiatry, Temple University Medical College.

[††] Assistant Professor of Psychiatry, David Geffen School of Medicine, University of California at Los Angeles.

[1] ANTHONY KENNY, ARISTOTLE'S THEORY OF THE WILL 11-48 (1979).

how people respond to incentives."[2] Yet it inevitably retains a normative implication as well, the implied contrast with the phenomenon of irrationality.

As a descriptive theory, RCT has come under attack from two disparate directions. People wary of reductionist science complain that it under-recognizes empathic transactions; they claim that it depicts as natural—and thus promulgates—a selfish, money-grubbing society.[3] At the same time, empirical researchers find that it fails to predict important examples of behavior exhibited by well-informed subjects in experiments within behavioral science.[4] Furthermore, the examples documented by systematic analysis are only a small proportion of the behavior patterns that people say they do not want but seem unable to give up. Seemingly free choice has led not only to alcoholism and drug abuse in a significant minority of people, but also to an epidemic of overeating, credit card abuse, overconsumption of passive entertainment, and other bad habits too widespread to be diagnosed as pathological.[5]

RCT arose not so much from empirical research as from a theoretical analysis of what decision strategies will dominate in market-

---

[2] Russell B. Korobkin & Thomas S. Ulen, *Law and Behavioral Science: Removing the Rationality Assumption from Law and Economics*, 88 CAL. L. REV. 1051, 1055 (2000). For further elaboration on rational choice theory, see Robert Sugden, *Rational Choice: A Survey of Contributions from Economics and Philosophy*, 101 ECON. J. 751 (1991)

[3] *See, e.g.*, JOHN DUPRÉ, HUMAN NATURE AND THE LIMITS OF SCIENCE 148 (2001) ("Not infrequently positive economics assumes that the real question is about maximizing wealth measured in monetary terms .... An obviously preferable goal would be something like standard of living...."); BARRY SCHWARTZ, THE BATTLE FOR HUMAN NATURE: SCIENCE, MORALITY AND MODERN LIFE 247-48 (1986) (arguing that the disciplines of economics, sociobiology, and behavior theory "have contributed to, and helped justify, the conditions that foster the pursuit of economic self-interest to the exclusion of almost all else").

[4] For an example in political science, see DONALD P. GREEN & IAN SHAPIRO, PATHOLOGIES OF RATIONAL CHOICE THEORY (1994). For a source regarding economics, see RICHARD H. THALER, QUASI RATIONAL ECONOMICS (1991); and for psychology, see Leonard Green et al., *Discounting of Delayed Rewards: A Life-Span Comparison*, 5 PSYCHOL. SCI. 33 (1994); Daniel Kahneman & Amos Tversky, *The Simulation Heuristic*, *in* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 201 (Daniel Kahneman et al. eds., 1982); George Loewenstein, *Out of Control: Visceral Influences on Behavior*, 65 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 272 (1996).

[5] *See generally* LAWRENCE J. HATTERER, THE PLEASURE ADDICTS (1980) (describing the process of addiction); ROBERT KUBEY & MIHALY CSIKSZENTMIHALYI, TELEVISION AND THE QUALITY OF LIFE: HOW VIEWING SHAPES EVERYDAY EXPERIENCE (1990) (examining television addiction); Avner Offer, Epidemics of Abundance: Overeating and Slimming in the USA and Britain Since the 1950s (Nov. 1998) (unpublished manuscript) (discussing the "social epidemics of overeating and slimming"), *available at* http://www.nuff.ox.ac.uk/economics/history/paper25/25offera4.pdf.

places.[6] To prevail over any significant period of time, an intention must be stable, and stability requires the standard properties of rationality—particularly commensurability, transitivity, and invariance across contexts.[7] However, this approach makes rational choice theory a set of rules for winning play—a normative model—rather than a description of how choice actually works. Even in the far simpler world of game theory, human subjects in experiments notoriously fail to follow obvious strategies that would increase their success.[8]

Psychologist Daniel Kahneman and his collaborators have directly studied human utility maximization. They estimated "objective happiness" by taking subjects' numeric self-reports of happiness moment by moment and calculating the area under the resulting curve over time.[9] According to the researchers, "[l]ogical analysis suggests" that the utility that a person derives from a particular event should be equal to the integral of all the instants that make up that event.[10] The integral that the researchers derived from this experiment did not reflect the way that the same subjects chose between the very experiences that they had evaluated this way. Subjects did not prefer those experiences with the greatest summed happiness (or least summed unhappiness), but displayed various perceptual distortions, particularly overvaluation of the greatest momentary reading and the latest reading.[11] These experiments demonstrate that real world utilities, as evaluated by RCT, fail tests for transitivity and invariance and, therefore, that conventional RCT is flawed as a descriptive theory.

In this Article, we present a utility-based model that fixes the major problems of RCT. We argue that decisions are determined in a single *intra*personal marketplace on the basis of a unitary selective principle—reward (Part I)—but the basic shape of the discount

---

[6] The pioneer was Paul A. Samuelson, *A Note on Measurement of Utility*, 4 REV. ECON. STUD. 155, 155 (1937) (presenting a theoretical method for measuring the marginal utility of income).

[7] *See* Korobkin & Ulen, *supra* note 2, at 1064 (stating that commensurability, transitivity, and invariance are some of the necessary conditions of rational behavior).

[8] *E.g.*, Richard H. Thaler, *The Ultimatum Game*, J. ECON. PERSP., Fall 1988, at 195, 196-98.

[9] Daniel Kahneman, *Objective Happiness*, *in* WELL-BEING: THE FOUNDATIONS OF HEDONIC PSYCHOLOGY 3, 5 (Daniel Kahneman et al. eds., 1999).

[10] *Id.* at 3.

[11] Charles A. Schreiber & Daniel Kahneman, *Determinants of the Remembered Utility of Aversive Sounds*, 129 J. EXPERIMENTAL PSYCHOL.: GEN. 27, 27-28 (2000) (discussing the results of experiments testing the peak-end rule).

curve for this reward creates conflicting interests within the individual (Part II). These conflicts can be partially resolved by perception of a prisoner's-dilemma-like relationship among successive motivational states, which generates will (Part III), but which does not approach rationality closely (Part IV). This model has practical implications for law and economics, the direction of which we can only suggest (Part V).

I. ALL REWARD MUST BE COMMENSURABLE

While RCT is unable to account for certain behavioral phenomena, in rejecting it there is a risk of throwing out the baby with the bath water. It is true that people do not behave so as to maximize attainment of a stable and ordered set of goals. Indeed, we will argue that they do not maximize any quantum without regard to the time at which they make their choice. Such a finding, however, does not require the conclusion—often encountered in cognitive psychology—that choice is irreducibly particularistic and not constrained to maximize anything. We do not have to abandon RCT's assumption that motivated behavior occurs within a single internal market, with a single currency of transaction. Indeed, research in both neurophysiology and behavioral psychology points to just such a market.

Over the last half-century neurophysiologists have located, with increasing precision, brain sites that control the selection of behaviors. From early on, it was known that animals would work to receive electrical stimulation in the medial forebrain bundle and nucleus accumbens.[12] Rats that can press a lever to get stimulation in these areas have been observed to press continuously for hours, until they become too weak to go on.[13] The same portions of their brains, which form part of the mesolimbic (midbrain) reward circuitry, were later shown to be those excited by cocaine[14] and by all other rewarding drugs that have been studied.[15] Furthermore, advances in brain-

---

[12] *See* James Olds, *"Reward" from Brain Stimulation in the Rat*, 122 SCIENCE 878, 878 (1955) (noting that rats with electrodes implanted into certain areas of their brains would seek to stimulate themselves by activating the electrodes); James Olds & Peter Milner, *Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain*, 47 J. COMP. & PHYSIOLOGICAL PSYCHOL. 419, 426 (1954) (same).

[13] Olds, *supra* note 12, at 878.

[14] *See* Hans C. Breiter et al., *Acute Effects of Cocaine on Human Brain Activity and Emotion*, 19 NEURON 591, 591 (1997) (summarizing that cocaine causes signal increases in the nucleus accumbens and other areas).

[15] *See* Gaetano Di Chiara & Assunita Imperato, *Drugs Abused by Humans Preferentially Increase Synaptic Dopamine Concentrations in the Mesolimbic System of Freely Moving Rats*, 85

imaging technology have made it possible to observe the reward process in normal human volunteers. These studies have shown that activity in mesolimbic reward circuitry accompanies even minor rewards such as winning a dollar,[16] receiving a squirt of pleasant tasting juice,[17] or viewing an attractive face.[18] Thus, there is an observable physical process that responds proportionately to known rewards.

It may turn out that localization of reward is more complex than current data suggest. However, commensurability of incentives is, in the end, a logical requirement of any theory of voluntary behavior; otherwise there would exist choices that could not be made.[19] As neurophysiologists Peter Shizgal and Kent Conover point out, the ultimate basis of choice must include a comprehensive marketplace of incentives.[20] They state that "[f]or orderly choice to be possible, the utility of all competing resources must be represented on a single, common dimension."[21] A model in which the activity of a quantitative selective mechanism determines value and hence choice need not require the maximization of total utility, but prevailing choices must be doing something over time that induces more of this selective process (hereinafter *reward*) than their rejected alternatives did.

The data from neurophysiology only add anatomic specificity to the vast literature on behavioral psychology, which has shown that choice is exquisitely sensitive to small changes in incentive.[22] Unlike

---

PROC. NAT'L ACAD. SCI. U.S. 5274, 5274 (1988) (providing evidence that all drugs abused by humans stimulate dopamine transmission in the limbic system).

[16] *See* Brian Knutson et al., *Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens*, 21 J. NEUROSCI. RC159, at 2-3 (2001), *at* http://www.jneurosci.org/cgi/reprint/21/16/RC159.pdf (indicating that +$1.00 and +$5.00 cues elicited happiness in human participants).

[17] *See* Gregory S. Berns et al., *Predictability Modulates Human Brain Response to Reward*, 21 J. NEUROSCI. 2793, 2797 (2001) (using juice and water to study human reward regions).

[18] *See* Knut K.W. Kampe et al., *Reward Value of Attractiveness and Gaze*, 413 NATURE 589, 589 (2001) (showing that brain activity increases when viewing an attractive face, especially when the face is directed toward the viewer).

[19] *See, e.g.*, GEORGE AINSLIE, PICOECONOMICS: THE STRATEGIC INTERACTION OF SUCCESSIVE MOTIVATIONAL STATES WITHIN THE PERSON 28-32 (1992) (demonstrating that a person's multiple centers of choice must compete on the same decisional dimension).

[20] Peter Shizgal & Kent Conover, *On the Neural Computation of Utility*, 5 CURRENT DIRECTIONS PSYCHOL. SCI. 37, 37-38 (1996).

[21] *Id.*

[22] *See* Richard J. Herrnstein, *Method and Theory in the Study of Avoidance*, 76 PSYCHOL. REV. 49, 67 (1969) (reporting how choice responds to incentives). *See generally* RICHARD J. HERRNSTEIN, THE MATCHING LAW: PAPERS IN PSYCHOLOGY AND ECONOMICS (Howard Rachlin & David I. Laibson eds., 1997) (presenting Herrnstein's

neurophysiology, behavioral psychology has extensively studied the effect of delay on the incentive value of reward. This study of delay has suggested a way to reconcile subjects' failure to maximize expected reward with the strict determination of choice by reward.

II. THE VALUE OF REWARD IS INVERSELY PROPORTIONAL TO DELAY

It has long been known that people discount the value of delayed goods, although some writers from Plato[23] to the Victorian economist Jevons[24] have called such discounting irrational. However, no one now thinks that rationality means maximization of reward without regard to timing—that $1001 a year from now is worth more than $1000 now. Thus, the concept of reward maximization has had to include some kind of discounting from the moment of expected delivery back to the moment of decision. Financial markets long ago established a norm for how this discounting should take place—the loss of a constant proportion of remaining value per unit of time, or *exponential* curve, which is the only function that will not lead to changes of relative valuation among goods at different delays as time passes. People adopted this curve to such an extent that, as utility theory took mathematical shape, this curve was assumed to depict not only the normatively rational discount rate but also the one that people follow spontaneously.

However, precise preference experiments in the last third of the twentieth century found a natural curve with roughly the same appearance—bowed upward to show smaller decrements as delays get longer—but with significantly different properties. This is the hyperbolic curve, which makes value inversely proportional to delay. A variant of behavioral psychologist Richard Herrnstein's matching law

---

theory of choice). Parametric motivational studies of children have shown both a continuity with the animal literature and the apparent effects of cultural overlay, which reduces the efficiency of some children in getting reward. *See* Edmund J.S. Sonuga-Barke et al., *Children's Choice: Sensitivity to Changes in Reinforcer Density*, 51 J. EXPERIMENTAL ANALYSIS BEHAV. 185, 196 (1989) (suggesting that internalized norms sometimes make children unresponsive to schedules of reward).

[23] *See* PLATO, PROTAGORAS 61-64 (Gregory Vlastos ed., Martin Ostwald trans., 1956) (discussing the choice between pleasures and pains).

[24] *See* W. STANLEY JEVONS, THE THEORY OF POLITICAL ECONOMY 70-74 (London, MacMillan 2d ed. 1879) (explaining the role of time on measurements of pleasurable effect).

as applied to delay,[25] this curve is adequately described by James Mazur's simple formula:[26]

$$Value = \frac{Value\ at\ No\ Delay}{1 + (Constant \times Delay)}$$

> The constant describes the subject's degree of impatience in the situation under study. By varying this one element, investigators have been able to produce substantially better fits to choices among delayed rewards than were possible with the exponential curves assumed by most utility theories.[27]

This curve gives preference a property that most people would call irrational—an innate tendency to switch from better-later goods to poorer-earlier goods simply as the earlier goods become imminently available (Figure 1). Such an innate instability of choice would seem maladaptive and hence unlikely to survive in evolution. However, people demonstrate it regularly when making single-shot choices in many different modalities of reward, including not only physical rewards like food[28] and relief from noxious noise,[29] but also money.[30] The only situations in which people do not seem to show hyperbolic discounting involve financial planning, where classical economics often makes successful predictions. Furthermore, the

---

[25] *See* Shin-Ho Chung & R.J. Herrnstein, *Choice and Delay of Reinforcement*, 10 J. EXPERIMENTAL ANALYSIS BEHAV. 67, 67 (1967) (validating the matching law for delayed reinforcement).

[26] James E. Mazur, *An Adjusting Procedure for Studying Delayed Reinforcement*, *in* 5 QUANTITATIVE ANALYSES OF BEHAVIOR 55, 58-59 (Michael L. Commons et al. eds., 1987).

[27] *See id.* at 71-72 (arguing that Mazur's formula provides a better explanation of test results than exponential curves).

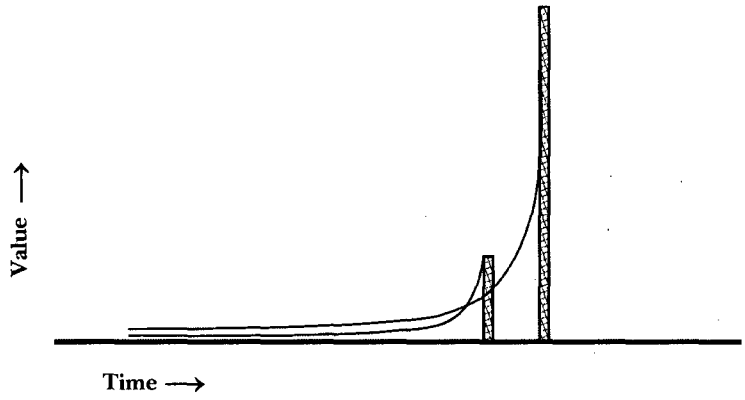
[28] *See* Steven P. Ragotzy et al., *Self-Control in Mentally Retarded Adolescents: Choice as a Function of Amount and Delay of Reinforcement*, 49 J. EXPERIMENTAL ANALYSIS BEHAV. 191, 195 (1988) (finding that mentally retarded adolescents changed preference from a smaller-earlier to a larger-later food reward as the delay before the earlier reward was increased).

[29] *See* Jay V. Solnick et al., *An Experimental Analysis of Impulsivity and Impulse Control in Humans*, 11 LEARNING & MOTIVATION 61, 74 (1980) (finding that a majority of participants in a study shifted preference from earlier, shorter terminations of irritating noise to delayed but longer terminations as the delay to the early terminations was increased).

[30] *See* Kris N. Kirby & R.J. Herrnstein, *Preference Reversals Due to Myopic Discounting of Delayed Reward*, 6 PSYCHOL. SCI. 83, 85-87 (1995) (noting that greater than ninety percent of participants in a study shifted preference from a smaller-earlier cash award to a larger-later one as delays to both awards increased); *see also* GEORGE AINSLIE, BREAKDOWN OF WILL 33-34 (2001) (arguing that most people would prefer $100 immediately to $200 in three years, but not $100 in six years to $200 in nine years).

universal observation of this pattern in animal experiments shows that this tendency is not an artifact of human culture, but an elementary property of a subject's response to reward.[31]

Figure 1



Hyperbolic discount curves from two rewards of different sizes available at different times (vertical hatched lines). The smaller-earlier reward is temporarily valued higher (preferred) for a period just before it is available, as shown by the portion of its curve that projects above that of the larger-later reward.

The finding of hyperbolic discounting at the root of reward valuation requires a radical reconceptualization of rationality and RCT. Accepting that reward is the selective factor for choice, and that people (and all behaving organisms) are constrained to maximize prospective discounted reward, makes the definition of rational choice both elementary and tricky. Rationality is clearly not maximization of actual reward at a single moment in time, which could best be accomplished by choosing tremendous short-term pleasure regardless of long-term consequences—perhaps smoking crack cocaine. The answer might be maximization of all discounted expected rewards—except that this criterion alone still includes, by definition, all behaviors that people actually choose, including smoking crack cocaine. Hyperbolic discounting implies that maximizing expected reward will sometimes entail not only preferring an objectively lesser good[32] over a greater one but also changing to this preference over time. Maximization of reward will seemingly dictate opposite choices when a smaller-

[31] *See* AINSLIE, *supra* note 19, at 63-76 (documenting animal and human experiments that demonstrate hyperbolic discounting).

[32] An objectively lesser good can be defined operationally as an alternative that is valued less when both options are available with no delay.

earlier good and a larger-later good are both distant and when the smaller-earlier good is imminent.

Thus, maximization per se can no longer be sufficient to define what would intuitively be called rational, since a person continually maximizing prospective reward would demonstrate radically unstable preferences that she herself, in the long view, would find unsatisfactory. Likewise, hyperbolic discounting offers a concept of irrationality that does not depend on a miscalculation of the contingencies of reward. Irrationality may simply be the person's choice of an alternative that she prefers only temporarily because of its proximity, rather than because she has misperceived its magnitude. By the same logic, two normative definitions of rationality may exist: rationality may be consistency, i.e., not undergoing temporary changes of preference, or rationality may be maximizing the longest-range goods, those that she would choose from the perspective of the greatest distance. These two possibilities are similar, but not identical, as we shall see. Either one permits a person to be irrational while still maximizing her discounted expected utility at every moment.[33] And in either case, this analysis separates the descriptive theory from the normative.

III. INVERSE PROPORTIONALITY PREDICTS A MECHANISM FOR WILL

We must leave the normative definition of rationality for a while and examine a more important question for RCT as a descriptive theory: What is reason and how does it contrast with passion? That is, what is the nature of the faculty that might make our choices consistent and/or maximize our longest-range good? It can no longer be just knowledge of the true contingencies of reward; a person can know through firsthand experience that drinking is not in her long-range interest and accordingly plan not to drink, but go on a binge when the opportunity arises. Some centuries after Plato, philosophy

[33] Neither definition requires rationality to call for selfishness, as anti-reductionists have feared. However, a discussion of this point must involve the consequences of hyperbolic discounting for emotion and empathy, which we cannot begin here. *See* AINSLIE, *supra* note 30, at 161-89 (discussing how hyperbolic discounting leads people to satiate appetites prematurely, including those for emotions; and in response, we look to vicarious experience and risk as "good source[s] of occasions for emotional reward"); *see also infra* note 88 and accompanying text (noting that "people deviate dramatically from rationality with respect to risk behavior").

found an additional faculty necessary to bridge the gap between insight and behavior: the will.[34]

A. *The Cognitive View of Will Is Hierarchical*

Cognitive psychology thinks of will as an executive at the top of a hierarchy, which is similar to the common-sense understanding of the concept. Elizabethan political philosophers, for instance, revealed their view of the self by their analogies of states to selves: there was a monarch on top (the brain) who gave orders to noble subordinates (voluntary muscles) who controlled the potentially rebellious masses (organs like the stomach and genitals).[35] States have since evolved to be more or less free clearinghouses of their citizens' wishes, in which leaders bid for votes and the winners negotiate with one another to deliver what they have bid. A chief executive is a convenience, often serving at the pleasure of one group of elected leaders or another; she survives by brokering the interests of these leaders. Leaders of even supposedly absolute command structures, like armies[36] or corporations,[37] are now recognized as unable to rule by fiat, but are obliged to deploy their influence so as not to lead their followers too far from where the followers want to go. However, the cognitive—and folk—models of the self are still monarchical or bureaucratic. Cognitive theorists have posited that self-control is literally an organ like a muscle, exhausted in the short run by use and strengthened in the long run by practice, but directed by an unspecified executive that presumably follows reason.[38] The agent of self-control (or "self-regulation") suppos-

---

[34] A will is still not universally held to be necessary for rational conduct. *See* GILBERT RYLE, THE CONCEPT OF MIND 68-69 (Univ. of Chi. Press 1984) (1949) (asserting that will is a superfluous concept); Gary S. Becker & Kevin M. Murphy, *A Theory of Rational Addiction*, 96 J. POL. ECON. 675, 685-92 (1988) (arguing that straightforward economic incentives can account for decisions not to consume addictive goods, without appealing to a self-control process).

[35] EUSTACE M. TILLYARD, THE ELIZABETHAN WORLD PICTURE 96-99 (1960).

[36] *See* Geoffrey Brennan & Gordon Tullock, *An Economic Theory of Military Tactics: Methodological Individualism at War*, 3 J. ECON. BEHAV. & ORG. 225, 226 (1982) ("Armies must be analyzed as collections of independent individuals who are, in some senses, as much at war with one another and their own leaders as they are with enemy forces.").

[37] *See* NILS BRUNSSON, THE IRRATIONAL ORGANIZATION 10-21 (1985) (arguing that there is no automatic link between a corporate decision maker's choice and organizational action).

[38] *See* Mark Muraven & Roy F. Baumeister, *Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle?*, 126 PSYCHOL. BULL. 247, 248 (2000) (asserting that "[t]he resource needed for self-control is a limited, consumable strength, much like a muscle's ability to work").

edly stands above motives and picks and chooses from them without being coerced by them:

> Misregulation occurs because [people] operate on the basis of false assumptions about themselves and about the world, because they try to control things that cannot be directly controlled, or because they give priority to emotions while neglecting more important and fundamental problems.[39]

According to cognitive theorists, "desires" such as emotions are only *a* reason to make a decision, and a rather disparaged reason at that, rather than *the* reason. Implied is a cognitive homunculus that evaluates desire together with a number of other motives for deciding and makes an autonomous choice.

B. *Neurophysiology Says Little of the Will*

Neurophysiological techniques demonstrate changes of cortical activity in muscle-control centers when a subject plans body movements[40] or even observes movements by another person.[41] These observations are sometimes described as correlates of will,[42] but this is will in the trivial sense of intentionality. This brain activity does involve the kind of temporal perspective that has been called "prospective memory"[43] and may involve the suppression of alternative action plans,[44] but so far, neurophysiology has told us almost nothing about how people make their choices consistent.

---

[39] Roy F. Baumeister & Todd F. Heatherton, *Self-Regulation Failure: An Overview*, 7 PSYCHOL. INQUIRY 1, 13 (1996).

[40] *See* Lüder Deecke & Wilfried Lang, *Generation of Movement-Related Potentials and Fields in the Supplementary Sensorimotor Area and the Primary Motor Area*, 70 ADVANCES NEUROLOGY 127, 127 (1996) ("Modern neurophysiologic techniques enable us to study changes in cortical activity in association with specific motor or cognitive functions."); Benjamin Libet, *Do We Have Free Will?*, 6 J. CONSCIOUSNESS STUD. 47, 49 (1999) ("The brain . . . begin[s] the volitional process in [a] voluntary act well before the activation of the muscle that produce[s] the movement.").

[41] *See* Wolfgang Prinz, *Perception and Action Planning*, 9 EUR. J. COGNITIVE PSYCHOL. 129, 129-54 (1997) (arguing that the brain reacts in a common manner to a perception of a neighbor's action and to a plan of action by the subject herself).

[42] *See* David H. Ingvar, *On Volition: A Neurophysiologically Oriented Essay*, 6 J. CONSCIOUSNESS STUD. 1, 2-4 (1999) (describing the neurophysiological aspects of woluntary movement).

[43] For a collection of articles discussing "prospective memory," see PROSPECTIVE MEMORY (Maria Brandimonte et al. eds., 1996).

[44] *See* C.D. Frith et al., *Willed Action and the Prefrontal Cortex in Man: A Study with PET*, 244 PROC.: BIOLOGICAL SCI. 241, 241-46 (1991) (studying brain activity for willed acts where subjects had to make a choice between actions in comparison with activity for routine actions).

C. *Parametric Behavioral Data Provide a Basis for Will*

Hyperbolic discounting suggests a different and more explicit explanation than the hierarchical model of what executive functions do. If contradictory rewards often select for incompatible behaviors, then executive functioning must be more than just a matter of finding out the sizes of available rewards and directing efforts toward them. Executive functioning means resolving the conflicts and inconsistencies that these learned processes generate. Hyperbolic curves per se can be expected to turn reward-seeking into a free-for-all in a population of successive, incompatible processes. Processes that are learned when one reward is dominant have to include means to actively undermine rival processes, which were learned when an incompatible reward was dominant.[45] The processes that are learned to obtain one reward—the *interest* in that reward[46]—must behave strategically toward the interest in a differently timed alternative in just the same way that interests contend with one another in a body politic. They must incorporate processes that forestall rival processes as they bid for acceptance in an internal marketplace. Consequently, the study of an individual's choice making and choice maintaining must resemble the study of interpersonal marketplaces, a micromicroeconomics or *pico-economics.*[47]

The simplest measure that can be taken against a competing interest is to alter the environment in order to create a commitment to a current preference. A current preference for eating in moderation can be secured by undergoing gastric bypass surgery or, less permanently, by checking into a "fat farm." Typically, though, the commitment is less definitive, retaining the possibility of reversal, but changing the contingencies to make the alternative interest less attractive to the future self. Buying only healthy food at the supermarket does not guarantee that you will not go on a late night junk food binge, but it

---

[45] Thus, they predict the multiple selves postulated by Richard Posner, *Are We One Self or Multiple Selves? Implications for Law and Public Policy*, 3 LEGAL THEORY 23, 24-25 (1997), and others before him, *see, e.g.*, THE MULTIPLE SELF (Jon Elster ed., 1986) (exploring theories of the individual as a collection of "several selves"). Posner doubts the theoretical necessity of hyperbolic curves themselves to account for preference reversal as a function of time, Richard A. Posner, *Rational Choice, Behavioral Economics, and the Law*, 50 STAN. L. REV. 1551, 1555-56 (1998), but his alternative explanation, information cost, would not account for the emergence of warring "present-oriented" versus "future-oriented" selves.

[46] *See* AINSLIE, *supra* note 30, at 42-44 ("[T]he mental operations selected for by a particular kind of reward [can be called] the person's 'interest' in that award.").

[47] *Id.* at 47.

adds the disincentive of having to go to a store when the urge strikes. Proclaiming to your friends that you will never eat meat again does not eliminate it as an option. Instead, it adds a new cost—that of losing face. David Laibson has suggested that a need for this kind of commitment explains people's otherwise unaccountable preference for relatively illiquid investments.[48] Similar behavior has been demonstrated in pigeons, suggesting that a conflict of enduring long- and short-range interests is an elementary consequence of hyperbolic discount curves: Some pigeons that consistently choose (through the peck of a key) a lesser but immediate food reward over a greater but delayed food reward also peck a key in advance of the choice when it eliminates the future availability of the lesser, immediate reward.[49] At the time the committing key is presented, the discounted value of the greater but delayed food reward is the more compelling alternative (see the portion of Figure 1 to the left of where the discount curves intersect). Two other committing methods are also straightforward: (1) keeping your attention off of temptations, either consciously[50] or in the Freudian defense mechanisms of suppression, repression, and denial; and (2) preparing your emotions, either consciously[51] or in the defense mechanisms of isolation and reversal of effect.

While strategic commitment might fit into the broad class of executive functions, it is not ordinarily labeled as "willpower." Indeed it could more reasonably be classified as a technique that eliminates the need for willpower. Understanding willpower requires an understanding of dynamic strategic interaction across successive motivational states. Writers since antiquity have related self-control to choosing according to principle, that is, choosing in categories containing a number of expectable choices rather than just the choice at hand. For example, Aristotle said that incontinence (*akrasia*) was the result of choosing according to "particulars" instead of "universals";[52] Kant

---

[48] David Laibson, *Golden Eggs and Hyperbolic Discounting*, 112 Q.J. ECON. 443, 443-45 (1997).

[49] George Ainslie, *Impulse Control in Pigeons*, 21 J. EXPERIMENTAL ANALYSIS BEHAV. 485, 488 (1974).

[50] *See, e.g.*, Jane Metcalfe & Walter Mischel, *A Hot/Cool-System Analysis of Delay of Gratification: Dynamics of Willpower*, 106 PSYCHOL. REV. 3, 3-19 (1999) (describing children's deliberate efforts at attention control).

[51] *See* Harriet Nerlove Mischel & Walter Mischel, *The Development of Children's Knowledge of Self-Control Strategies*, 54 CHILD DEV. 603, 603-19 (1983) (discussing children's awareness of needing "cool thoughts" to control themselves in "delay-of-gratification situations").

[52] 2 ARISTOTLE, *Nichomachean Ethics*, *in* THE COMPLETE WORKS OF ARISTOTLE 1729, 1811-13 (Jonathan Barnes ed., 1984).

said that the highest kind of decision making involved making all choices as if they defined universal rules;[53] the early psychologist Sully explained how "action becomes pervaded by principle" so that will is unified when "[p]articular actions . . . are viewed as members of a class of actions subserving one comprehensive end."[54]

These writers' fundamental insight is that you increase your self-control by choosing according to category rather than on a case-by-case basis. You may prefer to be a nonsmoker generally, even though, considering only an individual choice, you prefer to smoke. But just such an effect is predicted by hyperbolic discount curves. Although hyperbolae spike up sharply in the period just before a reward is due and are thus exquisitely sensitive to short delays, their tails are higher and more level than the tails of exponential curves at long delays. The relatively high tails of hyperbolic curves imply a potential for great increases in value if a number of expected future rewards are added together. Exponential curves keep declining relentlessly at a constant proportion of their remaining height for every unit of time that passes. Hyperbolic curves level off. The height of their tails means that curves from a series of alternative rewards, if bundled together, will favor the larger-later rewards increasingly as the series lengthens (Figure 2A).[55] By contrast, exponential curves do not predict increased tolerance for delay with summation of a series of choices (Figure 2B).

Recent experiments confirm a greater tolerance for delay with bundled rewards. Kirby and Guastello gave college students choices between smaller-earlier rewards—of money in one experiment, food in another—and larger but more delayed alternatives.[56] In one condition, the same choice was given five times, each time separated by a week. In another condition, a single choice was made on the first week between the two alternatives for all five weeks at once. As predicted from the summation of hyperbolically discounted rewards, preference for the larger-later alternative was increased in the condi-

---

[53] IMMANUEL KANT, RELIGION WITHIN THE LIMITS OF REASON ALONE 18-20 (Theodore M. Green & Hoyt H. Hudson trans., Harper Torchbooks 1960) (1934).

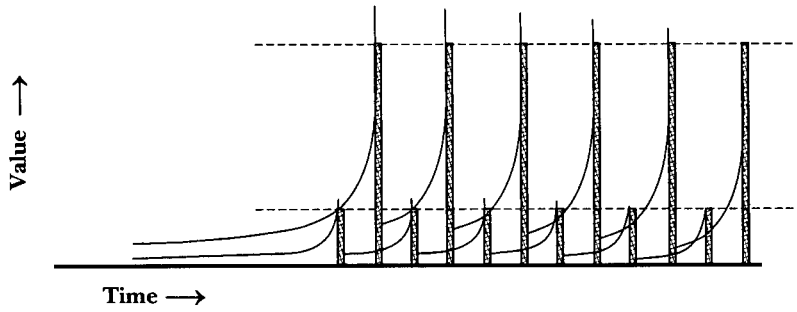
[54] JAMES SULLY, OUTLINES OF PSYCHOLOGY 631 (New York, D. Appleton & Co. 1891).

[55] There is behavioral evidence that the discounted effects of a series of rewards simply add. *E.g.*, James E. Mazur, *Choice, Delay, Probability, and Conditioned Reinforcement*, 25 ANIMAL LEARNING & BEHAV. 131, 141-43 (1997).

[56] Kris N. Kirby & Barbarose Guastello, *Making Choices in Anticipation of Similar Future Choices Can Increase Self-Control*, 7 J. EXPERIMENTAL PSYCHOL.: APPLIED 154, 154 (2001).

tion in which a series of choices was bundled together. We recently demonstrated the same phenomenon in rats, again showing that the implications of hyperbolic discounting do not depend on human culture. Eight rats were run through two conditions of a procedure designed to determine how many milliliters of sugar water immediately was equal in value to 150 milliliters after a three-second interval. In one condition of the procedure, rats made choices on a trial-by-trial basis, while in another condition, their choice determined the reward that they would receive for three consecutive trials. As predicted by hyperbolic discounting, all subjects tolerated more delay when the choices were bundled together.[57]

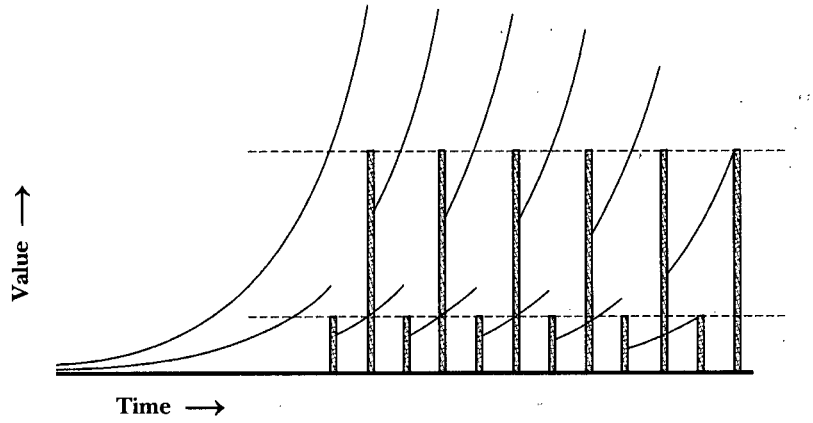
Figure 2A



Summed hyperbolic curves from a series of larger-later rewards and a series of smaller-earlier alternatives (vertical hatched lines). Each curve depicts the summed discounted values of all future (more to the right) rewards in the series. As the series gets longer and the summed curves peak higher above the current rewards, the initial period of temporary preference for the series of smaller rewards shrinks to virtually zero. (Compare the curves just before the first short vertical hatched line with that at the top of the the last short vertical hatched line.) The curves from the last (right-hand) pair of rewards are the same as in Figure 1.

[57] George Ainslie & John Monterosso, *Building Blocks of Self Control: Increased Tolerance for Delay with Bundled Rewards*, 79 J. EXPERIMENTAL ANALYSIS BEHAV. 37 (2003).

Figure 2B



Summed exponential curves from the same series of paired alternative rewards (vertical hatched lines). Summing increases their heights as the series get longer (more to the left), but does not change their relative heights. (This would also be true if the curves were so steep that the smaller-earlier rewards were preferred; but in that case summing would add little to their total height, anyway, because the tails of exponential curves are so low.)

D. *Intertemporal Bargaining Enforces Principles*

Thus, better, but more delayed, alternatives are more attractive when presented as a whole category of choices than they are individually; accordingly, a choice according to a general principle should favor such alternatives. However, a piece of the puzzle is still missing: If a person is a population of reward-seeking processes that have been learned wherever they are rewarded, what could make this throng choose according to principle? The need to explain principled choice is what usually makes theorists postulate innate executive processes—an organ like the ego. But what would empower an ego to apply a principle if it were powerless against the temporary preference to begin with? Such a theoretical organ fails to explain the selective power that even a hierarchical decision-making model needs, much less a marketplace model. Fortunately, hyperbolic discount curves suggest a way that the internal marketplace could, from its own basic properties, motivate the formation of the familiar executive processes.

If we imagine a Hobbesian state of nature within the individual, before the existence of an ego, then the life of any long-range plan will be short. Before it reaches its goal, an incompatible plan will become more attractive at some point. A child who wants friends may have too many urges to be selfish. Someone who wants to lose weight

may encounter too many tempting foods. An imminent payoff for an individual act of selfishness or particular snack is apt to be worth the little damage it does to friendships or the minor weight gain it causes. It would probably not be worth losing all expectation of friendship or slimness, but such huge outcomes are rarely the necessary consequence of individual choices. Therefore, in the state of nature, a person remains riddled with impulses. There is no incentive to plan because plans are usually rendered idle by the experience of reversing preferences.

However, an astute person—or someone who borrows the astuteness of her culture—is aware that her preferences are volatile. The best way she has to predict what she will do in the face of a future temptation is to see what she does with a similar temptation in the present: An act of selfishness predicts further selfishness and the eventual loss of friendship with all but the most long-suffering people. The snack predicts future snacks and inevitable weight gain. However, insofar as she is responsive to this rough insight about self-prediction, her current choices will become test cases—choices about selfishness and eating that this elementary insight will bundle together to form expected series of outcomes. When she chooses to be selfish, she chooses an expectation of future selfishness as well, and when she overeats, the act bodes more overeating. She will seem to be choosing according to principle, but what literally happens is that her successive selves form a repeated prisoner's dilemma relationship, which they come to solve in the same way as tacit interpersonal bargainers do. Each expects future selves to perceive the current choice as a precedent for cooperation or defection, and this expectation adds to those incentives that depend on that choice alone.[58]

Our hypothesis is that the will is an intertemporal bargaining situation, dependent for its force on a person's recursive evaluations of the prospects for her own behavior. Such an internally fed-back process is probably impossible to study with controlled experiments. However, it fits descriptions of will over the centuries better than other published theories of will and solves thought experiments that have otherwise seemed paradoxical in the philosophy of mind.[59]

---

[58] The terms of the prisoner's dilemma must be modified slightly to deal with the fact that future selves cannot retaliate strategically against past selves. *See* AINSLIE, *supra* note 30, at 92-93 (explaining that the prospect of defections by future selves, though motivated purely by their own future prospects, serves in practice as a threat of punishment).

[59] See *id.* at 126-29, 134-39, for discussions of Kavka's problem and Newcomb's

Thought experiments may prove to be a particularly useful way of isolating the active ingredient of subtle incentives like the value of precedents: Imagine that you are a smoker who is trying to quit but who craves a cigarette. Suppose that an angel appears to you and says that you are destined to smoke a pack a day beginning tomorrow. Given this certainty, you would have no incentive to turn down the desired cigarette today—it would seem pointless. What if the angel revealed instead that you were never to smoke again after today? Here, too, there seems to be little incentive to turn down the cigarette—it would be harmless. Fixing future smoking choices in either direction evidently makes smoking the dominant current choice over not smoking. Only if your future smoking is in doubt does a current abstention seem worth the effort. But its importance cannot come from any physical consequences for future choices—hence the conclusion that it matters as a precedent. Indeed when Kirby and Guastello merely suggested to student subjects that the subjects' current choices might serve as predictions of their future choices, preference for larger-later alternatives increased, although not as much as when the experimenters bundled the choices directly.[60]

The more explicit a person's perception that a current choice predicts a category of future ones, the more that perception will be true. A choice that you see as a test case will obviously carry more weight as an omen than a choice that is merely a random example, and a random example of a specific category will carry more weight than an isolated whim. Reiteration of this phenomenon can be expected to operate as a positive feedback system that increasingly distinguishes choices that are more systematic or principled from choices that are mostly spontaneous. According to our hypothesis, the mental processes involved in recognizing test cases and discerning their scope are the basic executive functions: Is selfishness forbidden even toward my rivals? Does a lapse in my diet reduce my will to not smoke as well? The shaping of the mental processes that answer such questions is based on the increased expectation of the long-range reward that such processes create. By brokering, in effect, large categories of expected reward (e.g., the aggregated expected value of the pleasures and suffering of being a smoker), they have the power to license and forbid behaviors. These executive functions, like the executives of parliamentary democracies, serve at the pleasure of the interests they gov-

problem. We deal below with the problem of free will, which can be seen as a similar thought experiment.

[60] Kirby & Guastello, *supra* note 56, at 160.

ern. The marketplace of motivated behavior has produced ego functions, much as Freud envisioned.[61]

E. *Recursive Self-Prediction Can Generate the Experience of Freedom of Will*

If future events were discounted exponentially, there would be no reason to make and enforce the personal rules by which you evaluate decisions as precedents. You could make a plan with the expectation that you would follow it, or, if circumstances change, modify it in a way that you currently have no reason to forestall. Your past and future selves would cooperate in an unbroken chain; their success in maximizing your good would be limited only by imperfect information or power. But with hyperbolic discounting you face a future that is chaotic, not only in the common meaning of the term, but also in its technical meaning: Your stream of future choices is *sensitively dependent* on your current one in that your current choice may send the stream in a wildly different direction than it seemed to be heading.[62] The greater your perception of how your choices matter as precedents, the greater this sensitive dependence.

Hyperbolic discounting makes decision making a crowd phenomenon, with the crowd consisting of the successive dispositions to choose that the individual has over time. At each moment you make the choice that looks best for you; but a big part of this picture is your expectation of how this choice will influence your choice at later times, an expectation that is mostly founded on the effects of your previous choices. Participation in the acts of this crowd of successive choice makers is a self-referential process, hidden from the outside observer and even from you yourself while facing it in advance. You can never be sure how your own future self will choose; you may read a small sign of faltering as your cue to bail out—that is, to stop cooperating with later selves on a given plan—just as investors may see a small drop in stock's price as a signal to start a massive sell-off. Or you may not. You will not know until it happens.

---

[61] *See* SIGMUND FREUD, FORMULATION ON THE TWO PRINCIPLES OF MENTAL FUNCTIONING (1911) ("[T]he substitution of the reality principle for the pleasure principle implies no deposing of the pleasure principle, but only a safeguarding of it."), *reprinted in* 12 THE STANDARD EDITION OF THE COMPLETE PSYCHOLOGICAL WORKS OF SIGMUND FREUD 213, 223 (James Strachey ed. & trans., Hogarth Press 1995) (1958).

[62] *See* Susan Ayers, *The Application of Chaos Theory to Psychology*, 7 THEORY & PSYCHOL. 373, 376 (1997) (explaining how "small changes in initial conditions [can] have large effects").

Intertemporal bargaining creates shifting trains of choices and consequences that could not be predicted from mere summation of the relevant motives. Behavior becomes like the weather—often predictable in the immediate future if you have a good knowledge of its driving forces, but subject to sudden shifts that are compounded recursively, making it unpredictable from a distance. If being unpredictable in principle is the necessary quality of a free will,[63] this bargaining may be what elicits free will from an underlying determinism.

Of course, mere dependency on internal feedback processes does not create the feeling of being a self:

> [I]f chaos-type data can be used to justify the existence of free will in humans, they can also be used to justify the existence of free will in chaotic pendulums, weather systems, leaf distributions, and mathematical equations.[64]

That is, even internal feedback processes, if they do not engage what feels like your self, will be experienced as random, "more like epileptic seizures than free, responsible choices."[65] We are arguing that intertemporal bargaining supplies that element of engagement: that your own motivation—in many cases emotion—is what you are predicting.

In conventional accounts, will is an irreducible process that does not—and indeed could not—predict itself:

> Making a decision and predicting that decision are mental states that exclude each other in the same mind, since making a decision implies, by the very meaning of the term, uncertainty as to what one is going to do.[66]

But hyperbolic discounting turns predicting a decision into an integral part of making that decision. Indeed, the only thing that differentiates making decisions from following whims becomes discernment of the self-referential consequences that are at stake (i.e., your expectations of your own future choices). Thus the prominence of the person's recursive intertemporal bargaining process reconciles determin-

---

[63] *See* DANIEL C. DENNETT, ELBOW ROOM: THE VARIETIES OF FREE WILL WORTH WANTING 151-52 (1984) (explaining how "chaotic systems are the source of the 'practical' . . . independence of things that shuffles the world").

[64] A.A. Sappington, *Recent Psychological Approaches to the Free Will Versus Determinism Issue*, 108 PSYCHOL. BULL. 19, 27 (1990).

[65] Robert Kane, *Two Kinds of Incompatibilism*, 50 PHIL. & PHENOMENOLOGICAL RES. 219, 231 (1989).

[66] Arthur Pap, *Determinism, Freedom, Moral Responsibility, and Causal Talk*, *in* DETERMINISM AND FREEDOM IN THE AGE OF MODERN SCIENCE 200, 201 (Sidney Hook ed., 1958).

ism with the experience of free will. Although clearly pulled by identifiable motives, a person's choice in such a process cannot be predicted with certainty, even by the person herself. Nevertheless, choice is as strictly determined as the weather.

IV. THE WILL DOES NOT NECESSARILY FOSTER RATIONALITY

Unfortunately, a person's perception of the intertemporal prisoner's dilemma relationship—and the willpower that results from this perception—cannot simply cure the problem of temporary preference. Willpower may be the best way we know to stabilize choice, but the intertemporal bargaining model predicts that it will also have serious side effects—side effects that have been observed by clinicians. Such bargaining does not let us estimate our best prospects from moment to moment as truly exponential discounting would. Rather, it formalizes internal conflict, making some self-control problems better, but some worse.

A. *The Will Maneuver Has Costly Side Effects*

These side effects need to be discussed. Where they are noticed at all, they are not recognized as the consequence of using willpower. In a dangerous split of awareness, people tend to see willpower as an unmixed blessing that bears no relation to abnormal symptoms such as loss of emotional immediacy, abandonment of control in particular areas of behavior, blindness toward one's own motives, or decreased responsiveness to subtle rewards. We argue that just these four distortions are to be expected to a greater or lesser extent from a reliance on personal rules. They may even go so far as to make a given person's willpower a net liability to her.

1. Rules Overshadow Goods-in-Themselves

The perception of a choice as a precedent often makes it much more important for its effect on future expectations than for the rewards that intrinsically depend on it. When this is true, your choices will become detached from their immediate outcomes and take on an aloof, legalistic quality. Ainslie has argued that this legalism underlies the self-control style that clinicians call compulsive.[67] It is a polar op-

[67] AINSLIE, *supra* note 30, at 205-25.

posite from impulsive temporary preferences, despite a usage that erratically equates them (e.g., "compulsive drinking").

It is often hard to guess how you will interpret a current choice when looking back on it. Did eating that sandwich violate your diet or not? Where your rules' criteria are ambiguous, cooperation with your future selves is apt to be both rigid and unstable. Unless you can find clear lines to use as boundaries, you may be uncertain as to whether, facing a choice in the future, you will look back at your current choice and judge it to have been a lapse. Under the influence of an imminent reward, you may claim an exception to a rule but later think you fooled yourself; that is, see yourself as having had a lapse. Conversely, you may be cautious beyond what your long-range interest requires out of fear that you will later see your choice as a lapse. This rationale exacerbates compulsiveness. Every lapse reduces your ability to follow a personal rule, and every observance reduces your ability not to. Errors in either direction impose costs that would never result from exponential curves, since those curves would not make choice depend on recursive self-prediction in the first place.

2. Rules Magnify Lapses

When you violate a personal rule, the cost is a fall in your prospect of getting the long-range rewards on which it was based. But this prospect is what you have been using to stake against the relevant impulses. A lapse suggests that your will is weak, a diagnosis that may act recursively to weaken your will. After weeks in which the expected value of a future without smoking was enough to win out over the immediate value of smoking, a moment comes in which the occasion to smoke is so attractive as to make it the more powerful of the incompatible interests. This one choice can affect the stream of future choices; for some time to come the prospect of future abstinence may not be sufficiently credible to offer much of a stake against the immediate reward of smoking. One lapse thus weakens the will, an outcome that has been called "the abstinence violation effect."[68]

To save your expectation of controlling yourself generally, you will be strongly motivated to find a line that excludes from your larger rule the kind of choice where your will failed. This means attributing the lapse to a particular aspect of your present situation, even though

---

[68] G. Alan Marlatt & Judith R. Gordon, *Determinants of Relapse: Implications for the Maintenance of Behavior Change*, *in* BEHAVIORAL MEDICINE: CHANGING HEALTH LIFESTYLES 410, 410-27 (Parle O. Davidson & Sheena M. Davidson eds., 1980).

it will make self-control much more difficult when that aspect is present in the future. You may decide that you cannot resist the urge to panic when speaking in public, or to lose your temper at incompetent clerks, or to stop a doughnut binge once begun. Your discrimination of this special area has a perverse effect, since within it you see only failure predicting further failure. If you no longer have the prospect that your rule will hold here, these urges will seem to command obedience automatically without an intervening moment of choice. Where such encapsulated impulses are clinically significant, they get called a symptom—for instance, a phobia, a dyscontrol, or a substance dependence.

Thus, the perception of repeated prisoner's dilemmas stabilizes not only long-range plans but lapses as well.[69] An alternative, cognitive model of self-control failure based on exhaustion of "strength"[70] does not account for regular failure that is specific to a particular circumstance.

3. Rules Motivate Misperception

Personal rules depend heavily on perception—noticing and remembering your choices, the circumstances in which you made them, and their similarity to the circumstances of other choices. And since personal rules organize great amounts of motivation, they naturally create temptations for you to suborn the perception process. When a lapse is occurring or has occurred, it will often be in both your long- and short-range interests not to recognize that fact. Your short-range interest is to keep the lapse from being detected so as not to invite attempts to stop it. Your long-range interest is also, at least partially, to keep the lapse from being detected. Acknowledging that a lapse has occurred would lower the expectation of self-control that you need to stake against future impulses.

After a lapse, the long-range interest is in the same awkward position as that of a country that previously threatened to go to war under a particular set of circumstances, which then materialized. The country wants to avoid war without destroying the credibility of its threat and may therefore look for ways to be seen as not having detected the circumstance. Your long-range interest will suffer if you catch yourself

---

[69] *See* AINSLIE, *supra* note 30, at 91-97 (discussing effects of prisoner's dilemmas).

[70] *See* Roy F. Baumeister & Todd F. Heatherton, *Self-Regulation Failure: An Overview*, 7 PSYCHOL. INQUIRY 1, 3 (1996) (hypothesizing that "strength" at a given moment is at the same level for all endeavors).

ignoring a lapse, but perhaps not if you can arrange to ignore it without catching yourself. This arrangement, too, must go undetected; therefore, a successful process of ignoring must be among the many mental expedients that arise by trial and error—the ones you keep simply because they make you feel better without realizing why. As a result, money disappears despite a strict budget, and people who "eat like a bird" mysteriously gain weight.

4. Rules May Serve Compulsions

The fact that a decision comes to be worth more as a precedent than it is worth in its own right does not necessarily imply that it is the wrong decision. On the contrary, you would think from the logic of summing discount curves that the evaluation of choices in whole categories rather than by themselves would have to improve your overall rate of reward (Figure 2A). Cooperation in a repetitive prisoner's dilemma would have to serve the players' long-range interests or else they would abandon it. How, then, can self-enforcing rules for intertemporal cooperation ever become prisons? Why should anyone ever conclude that she was trapped by her rules and even hire a psychotherapist to free herself from a "punitive superego"?

The most likely answer is that a person can discern many possible precedents in a given situation, and the way of grouping choices that finally inspires intertemporal cooperation need not be the most productive. This is because of the selective effect of distinctness: Personal rules operate most effectively on distinct, countable goals. A rule can be self-enforcing only if each criterion for having followed it yields an unambiguous either/or test. A rule to maximize a good will be effective only if the good can be clearly quantified; thus, the ease of comparing all financial transactions makes the money fluctuate less over time than, say, the value of an angry outburst or of a night's sleep. The motivational impact of a prospective series of moods has to be much less than that of an equally long series of cash purchases.

So, cooperation among successive motivational states does not necessarily bring the most reward in the long run. The mechanics of policing this cooperation may produce the intrapsychic equivalent of regimentation, which will increase your efficiency at reward getting in the categories you have defined, but reduce your sensitivity to less easily categorized kinds of reward.

B. *These Side Effects Cause Pure Will to Fall Short of Rationality*

The attempt to optimize your prospects with personal rules confronts you with the paradox of definition—that to define a concept is to alter it, in this case toward something more formalized. If you conclude that you should maximize money, you become a miser; if you rule that you should minimize your vulnerability to emotional influence, you will develop the numbing insensitivity that clinicians have named alexithymia;[71] if you conclude that you should minimize risk, you become obsessively careful; and so forth. The logic of rules may come to so overshadow your responsiveness to experience that your behavior becomes formal and inefficient. A miser's strict rules for thrift make her too rigid to optimize her chances in a competitive market, and even a daring financier undermines the productiveness of her capital if she rules that she must maximize each year's profit.[72] Similarly, strict autonomy means shielding yourself against exploitation by others' abilities to invoke your passions. But alexithymics cannot use the richest strategy available for maximizing emotional reward—the cultivation of human relationships.[73] Likewise, avoidance of danger at any cost is poor risk management.

In this way, a person who depends on willpower for impulse control is in danger of being coerced by logic that does not serve what she herself regards as her best interests. Concrete rules dominate subtle intuitions, and even though she has a sense that she will regret having sold out to them, she faces the immediate danger of succumbing to short-range urges, like addictions, if she does not. If she has not learned ways of categorizing choices that permit subtle criteria to hold their own against concrete tests for intertemporal cooperation, the implications of these tests will make her compulsive.

To summarize our hypothesis of the origin and nature of "ego functions": A person moves beyond the state of nature by discovering self-prediction, and thus adds strategic considerations to her incen-

---

[71] *See generally* John C. Nemiah, *Alexithymia: Theoretical Considerations*, 28 PSYCHOTHERAPY & PSYCHOSOMATICS 199 (1977) (discussing theoretical models applicable to alexithymia).

[72] *See* Ali R. Malekzadeh & Afsaneh Nahavandi, *Merger Mania: Who Wins? Who Loses?*, 8 J. BUS. STRATEGY 76, 79 (1987) (reporting that the discipline of always maximizing annual profit is counterproductive).

[73] *See* George Ainslie, *A Utility-Maximizing Mechanism for Vicarious Reward: Comments on Julian Simon's* "Interpersonal Allocation Continuous with Intertemporal Allocation," 7 RATIONALITY & SOC'Y 393, 393 (1995) (noting that the "richest source of external occasions to gamble on is the apparent experience of other people").

tives for choice. Bundling choices into categories of related precedents increases her long-range reward, but if she puts too much reliance on explicit personal rules, the four side effects of this reliance may do as much harm to her longest-range interest[74] as her original inconsistency did. It will be in her longest-range interest to learn processes beyond simple categorization, in order to balance the value of consistency against that of flexibility and spontaneity, and to balance the value of defining precedents against that of other ways to influence future selves. The need for such subtlety further shapes the executive faculty that must serve at the pleasure of its constituents—indeed, literally depends upon their pleasure. The self, as defined behaviorally, could be considered inborn only inasmuch as the hyperbolic discount function that creates the incentive to learn intertemporal bargaining skills is inborn.

C. *Overreliance on Will May Foster Addictions*

Modern culture has been slow to recognize the dilemma of personal rules: that we are endangered by our willpower as well as by our impulses. For instance, modern writers wring their hands about both the average citizen's rising body mass index and the prevalence of dieting in the young,[75] without noting the implication that the enemy is now approaching from two opposite directions.

In the interpersonal realm, the dangers of rules are much better known. The English long ago established courts of equity to correct distortions that arose from the rigidity of laws, and the great social rule maker Jeremy Bentham cautioned that rules should not be fully binding.[76] A recent review by Cass Sunstein makes it clear that social control by rules creates side effects analogous to our problems 1, 3, and 4: the need for preserving precedents makes rules too rigid; this rigidity "drive[s] discretion underground" into transactions that are

---

[74] We assume the longest-range interest to be a single, unitary interest, because the tails of hyperbolic curves are proportional to the objective—that is, nondelayed—values of rewards when the times to the alternatives are much greater than the times between them. Preference among distant alternatives should be stable as long as they remain distant.

[75] *See, e.g.*, RICHARD A. GORDON, ANOREXIA AND BULIMIA: ANATOMY OF A SOCIAL EPIDEMIC (1990) (noting the increasing pressure on young women to conform themselves to an unrealistic body-ideal); Cara B. Ebbeling et al., *Childhood Obesity: Public-Health Crisis, Common Sense Cure*, 360 LANCET 473, 473 (2002) (discussing the prevalence of childhood obesity).

[76] Cass R. Sunstein, *Problems with Rules*, 83 CAL. L. REV. 953, 1007 (1995) (citing GERALD J. POSTEMA, BENTHAM AND THE COMMON LAW TRADITION 418-21 (1986)).

not a matter of record; and the need to use available bright lines between what is and is not permissible both forbids innocuous activities and licenses cleverly defined harmful ones.[77] This last side effect is by no means confined to the realm of law. Quality assurance programs that focus doctors' motivation increasingly on measurable indicators of quality are reducing their clinical intuitiveness.[78] Problem 2, as well, is evident in the interpersonal sphere. For instance, some potential drug addicts may be saved by legal deterrence, but many who are not deterred become identified criminals and are worse off than they would be if drugs were not illegal.

The robustness of suboptimal rules may sometimes let addictions serve long-range interests. Better to be fat, you might think, than anorectic. Your will may become so confining that a pattern of regular lapses actually makes you better off in the long run. The lore of addictionology often attributes bingeing to a patient's inhibitedness in the rest of her life. General overcontrol is said to set up periodic episodes of breaking loose. The model of intertemporal bargaining predicted by hyperbolic discount curves[79] provides a specific rationale for this pattern. Rules that eliminate any large source of emotional reward will create a proportional motive for you to bypass or break those rules. If those rules have, in William James's phrase, "grown too narrow for the actual case,"[80] even your long-range interest will lie in partially escaping from them.

Thus, personal rules that become compulsions can create what are in effect alliances between long- and short-range interests. The person's occasional binge comes to serve as a corrective to the comparative sterility of such rules, a means of providing richer experiences than these rules allow, while its transient nature still limits the damage it does. The longest-range interest of an alcoholic who is too rigid when sober may be to tacitly foster the cycle of drunkenness and sobriety, rather than be continuously imprisoned by her rules.

Alcoholics are sometimes described as nicer, or more genuinely creative, or more fully human when drunk. Furthermore, some ad-

---

[77] *Id.* at 994 (discussing the overinclusive and underinclusive aspects of rules).

[78] Lawrence P. Casalino, *The Unintended Consequences of Measuring Quality on the Quality of Medical Care*, 341 NEW ENG. J. MED. 1147, 1148-49 (1999).

[79] *See supra* Part II (discussing the role of temporal delays in creating hyperbolic preference curves).

[80] Thomas Taffe, *Education of the Heart*, 45 CROSS CURRENTS 380, 383 (1995) (quoting WILLIAM JAMES, *The Moral Philosopher and Moral Life*, *in* THE WILL TO BELIEVE 184, 209 (Longman, Green & Co. 1927) (1896)).

dicts plan binges in advance. Such people may believe that their binges are undesirable—indeed, "rationality" will almost certainly dictate such a belief—but the therapists they hire find them mysteriously unresponsive to treatment. The patient who arranges for drinking several days in advance—who goes off the disulfiram that commits her to sickness if she drinks, for instance, or who brings bottles to her rehabilitation program for later use—cannot simply be yielding to a short-range impulse. This is behavioral evidence that she experiences her particular plan to give up drinking as a stricture which, even at a distance, appears to need hedging, although she may be unable to report any such thing.

The twin dangers of uncontrolled and overcontrolled behavior are vividly illustrated in the case of overeating. Most anorectics have started out by confronting a genuine eating problem[81] and have apparently discovered thereby the great sense of power that successful dieting confers. If they concentrate on maximizing their willpower, anorectics are apt to seriously reduce their spontaneity and impoverish their interpersonal relationships. Those who cannot tolerate the rigidity of strict anorexia may accept a pattern of binge-and-reform bulimia much like that of the binge drinker.[82]

Simply increasing the scope or severity of personal rules does not make behavior more rational, and almost no psychotherapies attempt it.[83] Most psychotherapy deals with problems of overcontrol—described by psychoanalysts as a punitive superego, by cognitive therapists as overgeneralization and magnification, and by gestalt therapists

---

[81] *See* Kunio Inanuma, *The Development of Anorexia Nervosa*, 40 JAPANESE J. CHILD & ADOLESCENT PSYCHIATRY 252, 252 (1999) (noting that dieting is sometimes a precursor to anorexia); Audrey R. Tyrka et al., *The Development of Disordered Eating: Correlates and Predictors of Eating Problems in the Context of Adolescence*, *in* HANDBOOK OF DEVELOPMENTAL PSYCHOPATHOLOGY 607, 616-17 (Arnold J. Sameroff et al. eds., 2d ed. 2000) (noting that "initial dieting" in response to "weight concerns" or "high body mass" are "significant predictors of subsequent eating disorders").

[82] *See* Janet Polivy & C. Peter Herman, *Etiology of Binge Eating: Psychological Mechanisms*, *in* BINGE EATING 173, 194-95 (Christopher G. Fairburn & Terence G. Wilson eds., 1993) (noting that restrictive dieting is often a precondition for the binge-purge cycle); Jane H. White, *Symptom Development in Bulimia Nervosa: A Comparison of Women with and Without a History of Anorexia Nervosa*, 14 ARCHIVES PSYCHIATRIC NURSING 81, 84 (2000) ("At least 28% of women with [bulimia] have a history of [anorexia] . . . .").

[83] An exception is WILLIAM GLASSER, REALITY THERAPY: A NEW APPROACH TO PSYCHIATRY 20 (1965) ("The more irresponsible the person, the more he has to learn about acceptable realistic behavior . . . .").

as dependence on cognitive maps[84]—rather than the simple inability to give up an impulse.

The tendency of overly concrete rules to keep your will from serving your longest-range interest is the great flaw of willpower. It suggests why a simplistic policy of "the more willpower, the better" contradicts the experience of many people with dyscontrol problems. To them, more willpower means less of the human qualities they value most in themselves. They are able to listen to reason only when reason, as represented by societal or personal rules, stops starving their own longest-range prospects for emotional satisfaction.

V. IMPLICATIONS FOR NORMATIVE RATIONALITY

For exponential discounters, consistency of choice would be a sign of unqualified success, of having estimated your long-range interest correctly from the beginning. For a hyperbolic discounter, however, consistency has a cost, sometimes an irrationally high one. Consistency per se may be foolish, indeed the hobgoblin of small minds. Although rules for consistency must produce reward in the relatively long range in order to survive, and may defend the person against impulses that dominate in the much shorter range, they may solidify behaviors that we would intuitively feel to be irrational; it is these behaviors that are experienced as compulsions. If people are hyperbolic discounters, they can either *impulsively* squander long range resources or *compulsively* imprison themselves for fear of their impulses while still strictly maximizing their expected discounted utility at every moment. Specifying optimal choice under these conditions is simple in theory but difficult in practice.

A. *Hyperbolic Discounting Demands a New Conception of Utility Maximization*

Given hyperbolic curves, a single theoretical definition of rationality stands out: the course of action that serves your longest-range interest. This will be the only course that is both intrinsically stable and has the intuitive advantage that—in the last analysis—you are glad to have followed it. However, even the long-range choice criterion does not provide a practical test for rationality. A person may or may not be able to assess which of her options will give her the most satisfac-

[84] These therapies are summarized in RAYMOND J. CORSINI, CURRENT PSYCHOTHERAPIES 4-17 (3d ed. 1984).

tion in retrospect;[85] but even if she could, that answer will not tell her beforehand which, if any, of her leading long-range prospects will survive the competition of faster-paying courses of action. She might be happiest in the long run if she buys an exercise bike and rides it regularly, but she will be least happy if she spends money on the bike and cannot marshal the motivation to ride it.

If tests of rationality are to provide prescriptions for choice, they will need to operate within the limits of sound game-theoretic strategy. This is to say that the behavioral economics of choice should not repeat the mistake of classical economics and evaluate options without regard to their strategic consequences. Sound macroeconomic choice evaluates the prospects of equilibria as determined by game theory,[86] and this must be true of picoeconomic choice as well. Evaluated strategically, rationality depends on which options can dominate other options for long enough to be realized. Judgments about which options should be avoided need to be supported by plans that use available incentives to make sure this avoidance is motivated.

In a recent review of behavioral economics, Christine Jolls, Cass Sunstein, and Richard Thaler classified the descriptive failures of RCT in three categories: bounded rationality, bounded willpower, and bounded self-interest.[87] Picoeconomics deals most directly with bounded (i.e., limited) willpower, as we have described. Much of bounded rationality seems to arise from pure cognitive error.[88] How-

---

[85] This limitation is the bounded rationality of Herbert Simon, which is familiar to conventional utility theory. *See* Herbert A. Simon, *Rational Choice and the Structure of the Environment*, 63 PSYCHOL. REV. 129, 136-38 (1956) (arguing that environmental and mental restrictions limit the degree of optimization that is achievable in real-life situations).

[86] *See generally* AMNON RAPOPORT, EXPERIMENTAL STUDIES OF INTERACTIVE DECISIONS, at ix (1990) (discussing the "interplay of theory and experimentation on group decision making in economics"); Vernon Smith, *Game Theory and Experimental Economics: Beginnings and Early Influences*, *in* TOWARD A HISTORY OF GAME THEORY 241, 244 (E. Roy Weintraub ed., 1992) (describing the shift toward the game-theoretic paradigm in economics).

[87] Christine Jolls et al., *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471, 1476-79 (1998).

[88] For example, people deviate dramatically from rationality with respect to risk behavior, in part because they do not incorporate base-rate probabilities in the manner delineated in Bayes's theorem. *See* Daniel Kahneman & Amos Tversky, *On the Psychology of Prediction*, 80 PSYCHOL. REV. 237, 257 (1973) (concluding that "[i]n making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction"). Another widely observed illusion of probability is described in P.C. Wason & J. Evans, *Dual Processes in Reasoning?*, 3 COGNITION 141, 148-52 (1975), in which the authors note that reliance on an individual's intuition often occurs at the expense of rational thought.

ever, some reported examples probably arise from strategic motives, either serving self-control (as when people pay a premium to keep money in an illiquid account)[89] or evading it (for instance, if the sunk-cost fallacy evades a personal rule for recognizing loss).[90] Hyperbolic discounting theory also provides a rationale for vicarious experience as a primary good, which can explain the apparent boundedness of self-interest.[91] Thus a correction for intertemporal bargaining might allow RCT to account in unified fashion for the greater part of the anomalies that have confronted it.

This prescription for rationality may include the personal-rule-governed consistency we have been discussing, but cannot depend on it exclusively. Rules that are too depriving are apt to fail, and this failure may either help or hinder your longest-range interest. Someone with a fear of her own spontaneity might find the greatest retrospective satisfaction from having faced her fear and renounced some of her rules for orthodox conduct. Realizing that she will probably not summon the courage to do this, is she rational to collude with the urge to go on binges, thereby enjoying some spontaneity but increasing her subsequent fear about how far it will go? It would be hard to say a priori. Ulysses will feel best in retrospect if he sails close to enjoy the Sirens' song and continues on, but he is rational in planning to do so only if he can reliably expect to be tied to his mast.

---

[89] *See* Christopher Harris & David Laibson, *Dynamic Choices of Hyperbolic Consumers*, 69 ECONOMETRICA 935, 937-38 (2001) (commenting that hyperbolic curves help to explain "pro-savings incentives like 401(k)s [and the] disproportionately low holds of liquid assists" among most individuals); *see also* AINSLIE, *supra* note 30, at 44 ("[H]yperbolic curves make a preference for illiquid savings rational—such savings serve as commitments.").

[90] *See* AINSLIE, *supra* note 19, at 291-93 (citing examples of individuals who avoid acknowledging a loss by counting it as part of a larger, still-viable gamble).

[91] A sketch of the argument: To show that empathic satisfactions can be treated like conventional economic goods, it is necessary to explain how emotional rewards, although available without fixed stimuli, are actually constrained by some kind of scarce condition. That scarce condition exists precisely because of hyperbolic discount curves: Maximal satisfaction from emotional rewards depends on their deferral and the consequent build-up of appetite for them; hyperbolic discount curves create a relentless urge to harvest these rewards prematurely. Therefore, unless people peg their emotions to occasions that are both optimally unresponsive to their current wishes and optimally surprising, their emotional lives will have the highly satiated quality of daydreams. The richest source of external occasions to gamble on is the apparent experience of other people, creating an incentive to "put ourselves in their shoes." AINSLIE, *supra* note 30, at 161-74; *see also* Ainslie, *supra* note 73, at 399 ("The best source of surprising, unique patterns is the behavior of other people.").

B. *The Law Will Be of Limited Help in Fostering Rational Choices*

An intertemporal bargaining model of impulsiveness and self-control is obviously relevant to the ultimate tool of social control, the law. However, its implication for action may not be what is apt to be the reader's first impression: that since people are not intrinsically rational, the law should add incentives to minimize temporary preferences for objectively poorer rewards. Hyperbolic discounting merely represents an experimental explanation of the well-known human trait once attributed to original sin—weakness of the flesh—and does not suggest that the solutions the centuries have provided are misguided. As we are about to argue, it only indirectly supports the controversial addition of another goal for criminal law—shoring up individual self-control, or "protecting you from yourself"—in addition to the conventionally accepted goals of deterrence, incarceration of the undeterrable, and gratification of people's wish for retribution. If a person is a population of partially conflicting interests that come to equilibria via a process like bargaining, supporting self-control is apt to be more than a matter of manipulating external incentives. A full analysis of legal implications is beyond our capabilities, but we should at least point out intertemporal bargaining theory's predictions of how an individual will respond to external incentives.

A population of interests that has been engendered by varied incentives is naturally divided, with some interests favoring both impulse and control sides of every major choice. This means that a person will be receptive to norms that might be useful criteria for personal rules against temptations, not just social rules to reduce conflict with neighbors. Granting Robert Scott's point that "behavioral science does not yet understand the mechanism of internalization [of norms],"[92] the demands of intertemporal bargaining (traditionally, "intrapsychic conflict") will certainly be an important incentive for this process; these demands will also prove to be an incentive to fit yourself into a social fabric. In other words, the conflict of short- and long-range interests will be a motivating factor in the process Robert Cooter has called "Pareto self-improvement"—the adoption of credible self-enforcing commitments.[93] Indeed, many behaviors that are illegal because they harm other people are also short-sighted from the

---

[92] Robert E. Scott, *The Limits of Behavioral Theories of Law and Social Norms*, 86 VA. L. REV. 1603, 1637 (2000).

[93] Robert Cooter, *Do Good Laws Make Good Citizens? An Economic Analysis of Internalized Norms*, 86 VA. L. REV. 1577, 1595 (2000).

viewpoint of individual self-interest. Even in the realm of crimes with victims, the criminal whose longest-range interest is to victimize other people is probably the exception rather than the rule; the attractiveness of most crime depends on the relative immediacy of its payoff.[94] This suggests that even when the law is protecting society from a person, it may be most effective by supporting the person's own longest-range interest in its competition with shorter-range interests, rather than threatening the person as a wholehearted wrongdoer. Here is a rationale for protecting the person from herself, but the process is tricky.

Authentication as "the Law" makes a norm stand out from other possible criteria for testing intertemporal cooperation, and thus gives it especial value.[95] However, the flip side of a person's need for external criteria is that she may feel vulnerable to manipulation; makers and enforcers of laws may seem to threaten her autonomy. If she does not see a law as serving her interests, the precedent she sets by obeying it may actually undermine her intertemporal cooperation. Likewise, even a benevolent authority who upsets a negotiated balance between the person's long- and short-range interests will undermine the intertemporal cooperation upon which they have compromised. Manipulation of incentives that benefits one interest too blatantly may have the same effect in a person as it does when an outside power takes a side in a civil war—to motivate middling factions to side with the opposition in order to preserve the person's, or country's, autonomy. Skillful intervention with a person, just as in Jerusalem or Northern Ireland, involves offering options that all sides find useful. This often includes providing ways to temper overly rigid rules.

External controls are also harmful in another way. According to the bargaining theory just presented, the phenomenon of will is generated by instances in which the person sees each choice as both necessary and sufficient to maintain intertemporal cooperation in an identifiable category and is thereby motivated to cooperate with her future selves. Whatever other incentives there are for cooperation, interventions that increase the number of relevant choices strengthen the will, while those that decrease the number of relevant choices

---

[94] *See* JAMES Q. WILSON & RICHARD J. HERRNSTEIN, CRIME AND HUMAN NATURE 49-56 (1985) (explaining the effect of time discounting on criminal behavior by using a metaphor of the tendency to choose to eat chocolate cake for the immediate sensory payoff rather than to wait for the long-term benefits of staying on a diet).

[95] For the role of bright lines in intertemporal bargaining, see AINSLIE, *supra* note 19, at 162-73.

weaken it. The incentive for maintaining personal rules is the expected value of all the better-later outcomes that hinge on obeying those rules. Where fewer outcomes hinge on a rule (e.g., because external incentives make the rule superfluous), the result will be a poorly motivated rule. A person may thus respond to the addition of external incentives by relaxing her own vigilance in the relevant domain. The personal rule "don't ever do *X*" is easily replaced by "don't do *X* when you might get caught." In an area where self-control had been robust and external policing is only marginally practical, this change might lead her to do more *X* rather than less. For example, when proctoring is added to the honor system in exams, it in effect replaces the existing system.[96]

Because of this perverse effect of coercion, rehabilitation programs that leave an alcoholic or credit abuser facing manageable temptations should be more effective than those that rely on external coercion to lock her up or to take her financial affairs out of her hands. Where no temptations are manageable at first, a treatment program should begin with a totally protected phase. This phase should be followed by a gradual return of the person's autonomy rather than a sudden confrontation with temptation, as often happens when a patient is discharged from an inpatient rehabilitation program. Any intervention that makes personal willpower less necessary will need to be followed by gradual exposure to temptations.

Even making committing devices available for a person's own use may not increase her self-control. It might seem, for instance, that a person should be able to commit herself contractually to follow through with a drug program, or lock in her most prudent future plans, and that courts should enforce these commitments even though they are unilateral. Some treatment programs have tried this on an informal basis;[97] but even a physical committing device like disulfiram (Antabuse), which makes alcohol sickening for a period of days, has been shown not to increase sobriety after one year in the absence of a highly structured social program.[98] Even when a physical

---

[96] Interference with personal rules may be a factor in how extrinsic incentives undermine people's autonomous decision making, an effect that has been widely reported. *E.g.*, EDWARD L. DECI & RICHARD M. RYAN, INTRINSIC MOTIVATION AND SELF-DETERMINATION IN HUMAN BEHAVIOR (1985).

[97] *See, e.g.*, Roger Paxton, *Deposit Contracts with Smokers: Varying Frequency and Amount of Repayments*, 19 BEHAV. RES. & THERAPY 117 (1981) (describing a program in which lapsing smokers forfeited money that they had deposited with the therapist).

[98] *See* Richard K. Fuller & Harold P. Roth, *Disulfiram for the Treatment of Alcoholism*, 90 ANNALS INTERNAL MED. 901, 904 (1979) (showing no statistically significant differ-

commitment scheme is self-enforced, it may suffer from the same drawbacks as one that is enforced externally.

A person's recognition of her susceptibility to both temptations and norms creates a strong motive for her to resist external influences, which in turn creates a problem for any efforts at social control. But even if the law could effectively encourage people to make their own personal rules stronger, it is not clear that more rational behavior would result, because of the four side effects we have described. Both the law and personal rules are rigid tools. Wielded with a heavy hand, they will interfere with the delicate art of rationality.

Where a person's behaviors are intolerable to others, the law must decide whether to preserve her will or ride roughshod over it. In the case where the person's behavior endangers others materially, the decision is usually forced; but where she is damaging herself and hurting others only through their empathic engagement with her, as in the case of drug abuse, intervention is of less clear value. Insofar as authorities want to protect her from herself, they need to consider the internal bargaining situation in which they will be intruding, including the possibility that ostensibly short-sighted behaviors are serving a strategic purpose for her longest-range interest.

C. *There Is No Natural Test for Criminal Responsibility*

Another controversial question in legal theory is when to hold a person criminally responsible when identifiable causes of her behavior are beyond her control. The law follows the popular intuition that it is fair to blame the person only when she could have done otherwise.[99] However, as science becomes increasingly proficient at detecting physiological and even genetic mechanisms for behavior, more and more misbehavior may seem to meet that test. Monterosso explored the basis of everyday notions of "could have done otherwise" and found that people gave physiological explanations the most weight by

---

ence between groups receiving disulfiram and a control group with respect to total abstinence, days worked, or family stability); *see also* C. Brewer, *Controlled Trials of Antabuse in Alcoholism: The Importance of Supervision and Adequate Dosage*, 86 ACTA PSYCHIATRICA SCANDINAVICA SUPPLEMENTUM 51, 51-55 (1992) (demonstrating the importance of close supervision to the successful use of Antabuse).

[99] *See, e.g.*, OLIVER WENDELL HOLMES, JR., THE COMMON LAW 54 (Little, Brown & Co. 1923) (1881) (noting that "it is felt to be impolitic and unjust to make a man answerable for harm, unless he might have chosen otherwise").

far.[100] The experiment described antecedent behaviors using vignettes that depicted individuals who engaged in undesirable behaviors of both a personal nature (e.g., overeating) and an interpersonal nature (e.g., violent temper episodes). The behavior was rated as significantly less voluntary and more excusable when its antecedent was specified as physiological (e.g., low levels of a particular neurotransmitter) as opposed to experiential (e.g., severe parental abuse). Physiological antecedents led to more judgments that the behavior was involuntary and thus less worthy of blame, as well as to the endorsement of more favorable treatment (lesser prison sentence or more healthcare coverage) for the individual described.[101]

Classifying behaviors as involuntary based on the presence of a physiological antecedent could eventually bring to full fruition the old maxim that "to understand all is to forgive all," and thereby undermine criminal deterrence generally. Subjects in the above study were willing to declare those behaviors that had physiological antecedents to be involuntary and inappropriate targets for punishment, despite having explicit information that the misbehaving individuals could weigh the consequences. The understanding of freedom of will proposed in Section III.E nevertheless permits a concept of criminal responsibility that is consistent with absolute determinism by assessing whether there was a fighting chance for available incentives, including the power of the law itself, to deter the behavior. Specifically, a person would be held responsible insofar as the prospect of being held responsible could realistically have played a role in her choices.[102]

The existence of a "disease," even a disease of motivation that, say, made alcohol abnormally alluring, would not in itself rebut responsibility. The attraction of alcohol may be part of a disease, just as an in-

---

[100] John Monterosso, Explaining Away Behavior: Scientific Analysis and the Transformation of Acts into Occurrences (1997) (unpublished Ph.D. dissertation, University of Pennsylvania) (on file with the University of Pennsylvania Library).

[101] For a similar effect in the domain of learning disabilities, see John Sabini & John Monterosso, *Moralization of College Grading: Performance, Effort, and Moral Worth*, 25 BASIC & APPLIED SOC. PSYCHOL. (forthcoming 2003).

[102] The common understanding of responsibility does seem to be a fairly constant perception despite wildly differing jury instructions, if experience with the insanity defense is good evidence. *See* Norman J. Finkel, *The Insanity Defense Reform Act of 1984: Much Ado About Nothing*, 7 BEHAV. SCI. & L. 403, 411 (1989) (showing that there was no significant difference in jury verdicts resulting from differing jury instructions on the insanity defense in a controlled experiment); Margaret A. McGreevy et al., *The Negligible Effects of California's 1982 Reform of the Insanity Defense Test*, 148 AM. J. PSYCHIATRY 744, 748-49 (1991) (finding that California's revision of the test for insanity had no practical impact on the rate of acquittal).

tense itch is sometimes part of a disease. The decision to gratify the urge is still subject to free will as we have defined it. Nevertheless, there are urges that few people can resist, however strong their self-control. Evidence exists that many seemingly reflexive behaviors are actually mediated by motivation and thus are ultimately choices. These behaviors include even the panic response to phobic stimuli and the emotion-like ("protopathic") component of pain that makes it aversive.[103] The fact that people can be taught not to respond to urges for these behaviors—for instance, in certain therapies for the pain of dental drilling or childbirth[104]—demonstrates that pain and panic are not reflexive but choices. However, resistance is so hard to learn that the law has to excuse its failure.[105] Likewise, a drug addict may have so damaged the credibility of her will that there was no realistic chance of her turning down a fix—and again the law might reasonably excuse her.

The presence or absence of disease has been a handy line to divide those behaviors that respond adequately to incentives from those that do not. Increasingly sensitive physiological measurement is obscuring this line. There is a continuum of prospects for resisting temptation ranging from the urge to panic to the casual wish for another dessert. Nature does not draw a line across this continuum at some point to serve the law's need for a dichotomy. However, the decision to excuse someone could be simply a recognition of her genuinely profound defeat, as bankruptcy is for a debtor—serving a practical purpose, so as not to waste resources for deterrence, as well as to feel fair. Such a recognition might be accompanied by legal disabilities that would protect others from the bad consequences of her defective will, just as if she were recognized to have a disease. This approach has the further advantage that it spares medical science the exercise of discerning the presence of supposed diseases of will in legal cases.

---

[103] *See* R. Melzack & K.L. Casey, *The Affective Dimension of Pain*, *in* FEELINGS AND EMOTIONS 55, 57-62 (Magda B. Arnold ed., 1970) (examining the motivational aspects of pain and sensory experience).

[104] For panic, see George A. Clum, *Psychological Interventions vs. Drugs in the Treatment of Panic*, 20 BEHAV. THERAPY 429 (1989); for pain, see J.C.R. Licklider, *On Psychophysiological Models*, *in* SENSORY COMMUNICATION 49, 50-51 (Walter A. Rosenblith ed., 1959); R. Melzack et al., *Stratagems for Controlling Pain: Contributions of Auditory Stimulation and Suggestion*, 8 EXPERIMENTAL NEUROLOGY 239, 245-46 (1963) (suggesting that structured attention can ameliorate pain perceptions based on experiments with auditory stimulation and other competing sensory stimuli).

[105] This holds true even in the case where succumbing to the urge harms others, as when panic causes an accident.

The likelihood that a person is a population of partially conflicting interests makes it easier to understand irrationality. However, it confirms what our culture has really known all along: that irrationality is not just a collection of errors but a robustly motivated phenomenon, and that correcting irrationality is a tenuous art at best. Controlled research on the resulting interactions has only begun and ultimately will be limited by their recursive nature, but the in the future any behavioral model will need to take them into account.