Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

EUROPEAN DOCTORATE IN LAW AND ECONOMICS

Ciclo __29_____

Settore Concorsuale: __13/A3_____

Settore Scientifico Disciplinare:___SECS-P/03_____

ESSAYS IN LAW & ECONOMICS

Presentata da: DANIEL YOONBUM PI

Coordinatore Dottorato

Prof. Luigi Alberto Franzoni

Supervisore

Prof. Francesco Parisi

Esame finale anno 2019

Essays in Law & Economics

Doctoral Dissertation

Daniel Pi[1]

December 5, 2018

[1]Visiting Assistant Professor, George Mason University. Contact: dpi@gmu.edu.
Thanks to Luigi Franzoni, Alan D. Miller, Francesco Parisi, and David Pi for their invaluable discussions and guidance.

Contents

1 Introduction **5**

2 Meta-Rational Choice **7**

2.1 Background 9

2.1.1 Expected Utility Theory 11

2.1.2 The Menagerie of Biases and Heuristics 14

2.1.3 Remedial Theories 19

2.2 A Meta-Model of Meta-Rational Choice 30

2.2.1 Notation 31

2.2.2 Actions and States of the World 32

2.2.3 Hierarchy of Decision-Making Procedures 35

2.2.4 Meta-Utility 39

2.3 Interpretation and Implications 45

2.3.1 Informal Restatement 45

2.3.2 Implications 49

2.3.3 Strong and Weak Interpretations of Meta-Rationality 52

2.3.4 Extensions 54

2.4 Conclusion 59

3 Bounded Criminality **65**

3.1 Background 66

3.1.1 The Elements of Criminal Deterrence 66

3.1.2 Criminological Skepticism about Economic Models of Deterrence 67

3.1.3 Becker's Approach to Crime 74

3.2 A New Framework for Deterrence 75

3.2.1 Bounded Rationality 76

3.2.2 An Example 78

3.2.3 Cashing Out the Public Policy 79

3.3 The Model . . . 81
3.4 Practical Considerations . . . 86
3.5 Conclusion . . . 89
3.6 Appendix . . . 90

4 Harmful Speech **93**
4.1 The Argument *for* Harm-Based Regulation . . . 94
4.1.1 General Theories of Speech Rights . . . 99
4.1.2 The Structure of First Amendment Review . . . 104
4.1.3 Defamation . . . 106
4.1.4 Espionage Act Cases and Incitement . . . 110
4.1.5 Fighting Words and Hate Speech . . . 114
4.1.6 Commercial Speech . . . 121
4.1.7 Obscenity . . . 125
4.2 Refutation of the Harm-Based Argument . . . 131
4.2.1 Endogenizing the *Harm* of "Harmful Speech" . . . 131
4.2.2 Endogenizing the *Benefit* of Harmful Speech . . . 133
4.2.3 Filtering . . . 135
4.2.4 Counter-Argument in General Form . . . 138
4.2.5 Counter-Counterarguments Anticipated . . . 139
4.3 Applications . . . 141
4.3.1 Affronting Speech . . . 142
4.3.2 Persuasive Speech . . . 153
4.4 Conclusion . . . 160

Chapter 1

Introduction

This dissertation collects two recent articles I have written and one older article. The first two articles develop a notion of tiered rationality. The first is a purely theoretical piece, and the second is an application of the concept to criminal law. The third article is distinct and separate from the previous two. It concerns countervailing effects in the regulation of speech.

The first article, "Meta-Rational Choice," addresses several foundational problems in the economic conception of the rational actor. The principal result is that for any possible decision-maker, there exists a real-valued function such that every choice the decision-maker would select could be represented as maximizing that function. The formal framework constructed to prove this result resolves several paradoxes associated with earlier conceptions of economic rationality. Several implications and interpretations of the result are explored.

The second article, "Bounded Criminality," applies the meta-rational approach to the analysis of criminal law. Law and Economics scholars have traditionally modeled criminal deterrence as a simple function two factors: (i) severity of punishment, and (ii) probability of punishment. Building on this insight, Gary Becker's (1968) seminal essay argued that optimizing on the benefits of deterrence and the cost of enforcement should be the policy objective of the criminal law. Yet empirical and experimental research in behavioral economics has shown that individuals often do not respond rationally to incentives. Instead, theories of bounded rationality predict that individuals form heuristics to guide their decision-making when decision costs are sufficiently high. This paper combines the insights of Becker's model of criminal deterrence with a theory of decision costs, arguing for a nuanced revision of the objective of criminal law: that the function of the criminal law should not be to create first order disincentives to commit particular crimes, but rather to instill

in individuals the formation of second order heuristics not to undertake a rational deliberation of the costs and benefits of criminal activity in the first place.

The final article, "Harmful Speech," presents an economic argument opposing the regulation of speech. Among the most common justifications for the regulation of speech is that some categories of expressive activity cause "harm." Conventionally, opponents of speech regulation have contested the putative "harmfulness" of supposedly harmful speech, or argued that the benefits outweigh the harms. This paper accepts, *arguendo*, the "harmfulness" of some classes of speech, but identifies three countervailing effects which suggest that regulation might be ineffectual, or that it may even exacerbate the harmfulness.

Chapter 2

Meta-Rational Choice

The engine for nearly all theoretical inquiry in economics is the hypothesis that individuals behave rationally. Its capacity to generate precise conclusions and adaptability to novel applications have allowed economics to envelope practically the whole of social science within its conceptual framework.[1] Yet it faces ever-growing skepticism. Over the past several decades, the assumptions upon which the hypothesis relies have suffered trenchant attacks,[2] and a mounting body of experimental work has revealed multifarious confounding observations.[3]

Real world behavior seems, at least in some circumstances, to deviate systematically from the rational actor hypothesis. These observations have elicited divergent responses from scholars. Conventionalists regard research undermining the rational actor hypothesis to be an assault on economic science, framing its findings as inconvenient obstacles rather than progress.[4] Behavioralists meanwhile celebrate the

[1]The expansion of economic methodology to non-economic subject matter is sometimes described by the appellative, "economic imperialism." Becker, 1976 gives an extended discussion of the cross-disciplinary applicability of economic theory. See also Hirshleifer, 1985, and Lazear, 2000.

[2]Allais, 1953; Ellsberg, 1961.

[3]This literature is vast. Some prominent examples include Camerer and Thaler, 1995; Kahneman, Knetsch, and Thaler, 1991; Tversky and Kahneman, 1981. For a general overview on the development of the field, see Baron, 2000; Thaler, 2015; Tversky and Kahneman, 1974.

[4]Defenders of the rational actor hypothesis have principally taken three distinct tacks. The first is to present evidence that real-world behavior does in fact tend to correspond with the predictions of the rationality assumption in most cases. See List, 2004 (finding that apparent deviations from the rationality assumption tend to dissipate as decision-makers acquire experience), The second is to dispute the methodological soundness of the countervailing evidence. See Plott and Zeiler, 2007, Wright and Ginsburg, 2012. Both approaches have produced interesting and persuasive results, and I am inclined to think they rightly tamp down the enthusiasm of overzealous behavioralists, who imprudently declare economic rationality to be a "refuted" hypothesis. Nevertheless, these efforts

development toward a more empirically grounded conception of decision-making.[5] And between these extrema there have arisen a myriad of nuanced intermediate positions.

Nearly all scholars recognize the pressing and fundamental nature of the questions posed. Even the most stalwart conventionalist must concede some answer is wanted. Yet attempts to amend or displace the "rational choice" paradigm with better foundations have failed to win widespread adoption. Thus, incongruously in the face of empirical challenges, the rational actor hypothesis remains the *modus operandum* of mainstream theoretical economics.

The field is presently in a state of skeptical suspension. Challenges to rational choice theory go unanswered, but no generally accepted alternative has yet emerged. We find ourselves inhabiting a precarious moment, which seems to possess all the telltale signs of an impending "scientific revolution,"[6] yet the revolution does not come.

Whether rational choice theory can be salvaged or whether we ought demolish the edifice to build upon wholly new foundations is a question of some moment. Yet we are not forced to choose between the two, for the alternatives are not mutually exclusive. In the face of controverting empirical evidence, some demolition is well warranted. However, what is built in its place can be made to resemble in its essence that which preceded it.

The contribution I aim to make is a modest one: the components of my theory are present already in the prior literature. The innovation proposed in this article is not to introduce new elements for consideration, but rather to arrange the puzzle

---

to preserve the orthodoxy seem indisputably to fall short of ruling out the existence of systematic non-rational phenomena altogether. And inasmuch as I am inclined to agree that the rational actor hypothesis cannot be so easily dismissed, neither do I believe the countervailing evidence can.

[5]For example, Korobkin and Ulen, 2000.

[6]Kuhn, 1962 claims that the condition for a "crisis" (potentially leading to a "paradigm shift" or "scientific revolution") is the discovery of experimental anomalies, which cannot be explained by prevailing scientific theory. This surely describes the present state of economics. By Kuhn's lights, the resolution of such anomalies is not necessarily the adoption of a new theoretical "paradigm," but in some (indeed most) cases, an extension of the existing paradigm. The question for economics is whether the present anomalies can be accounted for by "normal science," or whether their resolution requires a new theoretical basis.

It bears remarking that the applicability of Kuhn's model to "progress" in the social sciences is rather more controversial than its applicability in the natural sciences. And indeed, even with respect to the natural sciences, several of Kuhn's premises (most notably the incommensurability of paradigms) have been forcefully rebutted (for example, in Field, 1973). Regardless, while it is intuitively appealing to regard the present state of economics in Kuhnian terms, nothing critical to this article depends upon a Kuhn's conception of scientific progress.

pieces already on the table.

The meta-model presented in this article extends classical models of rationality to accommodate the presence of *decision costs*,[7] and *decision-making procedures*. Very many theories which could be construed as attempting something similar to my approach exist already. However, these earlier efforts have encountered profound conceptual and technical difficulties, in consequence of which they can hardly be considered "improvements" on conventional models of rationality. The models I present resolve the most critical problems which have heretofore plagued earlier attempts to formalize a general theory of behavior (along with some problems that have not received adequate scholarly attention) while retaining most of the advantages possessed by conventional theories of rationality.

2.1 Background

Let us start with some definitions. First, let "potential behavior" denote the counterfactual acts that an individual *would choose* given some state of the world. Next, let "utility functions" denote real-valued function meant to represent a decision-maker's preferences. And let "utility-maximization" refer to choice selections which correspond to the maximization of a utility function.

The "rational actor hypothesis" is the proposition that the potential choices of any decision-maker can be represented by utility-maximization. Given a set of available choices, the rational decision-maker will select the alternative which returns the greatest utility value. The rational actor hypothesis is the central conceit of "rational choice theory," which develops the hypothesis into a framework for analyzing incentives and behaviors in variegated circumstances and under varying constraints.

Interpreting the theoretical claim is not straightforward. The rational actor hypothesis may be understood as positing at least two distinct propositions. For clarity, let us disambiguate between a "weak" rational actor hypothesis and a "strong" rational actor hypothesis.

The weak rational actor hypothesis makes no *psychological* claim about decision-makers' deliberative processes. It holds merely that *whatever* cognitive mechanisms a decision-maker might employ in the formation of a choice, his actions are ultimately *representable* as the maximization of some utility function. The weak hypothesis critically denies (or at least reserves affirmation of) the "reality" of utility. Utility functions are—under the weak interpretation—merely "useful fictions," which

[7] This is a core claim of any theory of "bounded rationality." See generally, Gigerenzer, 2001; Gigerenzer and Goldstein, 1996; Simon, 1955.

impose a convenient structure upon reality.

By contrast, the strong rational actor hypothesis asserts that utility-maximization does not merely represent how individuals decide, but that it *is* how individuals decide.[8] I suppose rather few economists (and fewer decision theorists) would adopt the strong interpretation presently.[9] Nevertheless, there are compelling arguments for a strong interpretation, which I investigate further in this article.

The interpretive distinction is an important one. It is therefore worth a brief digression to develop a clear intuition on the difference between the weak and strong interpretations. Consider the following thought experiment: Imagine we set out to create robots capable of mimicking human decision-making. Suppose we construct a machine with all the external characteristics of an ordinary human being. It is physically capable of producing any behavior that a human being is physically capable of producing, and it is capable of accepting all the sensory inputs—information about its environment—that a human being is capable of perceiving. Internally, let us assume that our hypothetical robots are unconstrained by computational limitations.

Yet though they are *physically* capable of any actions that a human is, and though they "perceive" as well as humans do, and though they are unburdened by computational limitations, this is still insufficient to produce the behaviors that a human being might produce. We have not yet coded any instructions for the robot to execute. Suppose then that our hypothetical robots are designed to accept an initializing input when they leave the factory. Our robots are loaded with a "utility function"—an infinite array mapping all possible states of the world to points along a real number line. In any circumstance the robot encounters thereafter, it will select the action which maximizes the expected value of its utility function.

The weak rational actor hypothesis is equivalent to the claim: that for any given human being, there exists some utility function, such that if we loaded that utility function into a robot doppelgänger, then the robot's behavior would be indistinguishable from the human being's behavior (how we might *discover* the utility function which produces this effect is a separate issue—the point is merely that it *exists*).

The strong rational actor hypothesis is equivalent to the claim that all human beings simply *are* utility-maximizing robots (albeit composed of meat and bones

[8] I am inclined to believe that greater emphasis on this distinction would resolve many of the more trivial disagreements about the validity of the rational actor hypothesis.

[9] The distinction is sometimes described as differing in what is taken to be the atomic unit of analysis. What I refer to as the "strong rational actor hypothesis" may be construed as taking *preferences* as atomic, whereas the "weak rational actor hypothesis" takes *choice behavior* as atomic. See Mas-Colell, Whinston, and Green, 1995, p. 5. Some economists have distinguished between the weak hypothesis as relating to "decision utility" and the strong hypothesis as relating to "experienced utility."

rather than gears or circuitboards).[10]

Both interpretations assert that human behavior *can be represented* as mechanistic utility-maximization. They diverge as to the significance of the claim. The weak interpretation is agnostic as to the question *why* human behavior can be mimicked by utility-maximizing robots. The weak interpretation merely asserts that it *can* be. The strong interpretation commits to the proposition that human behavior can be mimicked by utility-maximizing robots, *a fortiori*, because humans simply *are* utility-maximizing robots.

2.1.1 Expected Utility Theory

The standard specification of the rational actor hypothesis is given by "expected utility theory," formalized in von Neumann and Morgenstern (1953).[11] Given that an individual's preferences are logically consistent, and assuming his choices reveal his preferences (the "revealed preferences hypothesis")[12] the von Neumann-Morgenstern

[10]This recasting of the rational actor hypothesis may clarify the intuition why the strong interpretation might be regarded as more extravagant. It should also provide some intuitive purchase as to my framing of these conceptions as differences in *interpretation*.

[11]The model is constructed axiomatically (note that the use of the term "axiom" is somewhat misleading. It is however the term that von Neumann, Morgenstern, and their successors used to denote their theoretical premises. Economists in the first half of the twentieth century seem to have been influenced by developments in logic and foundational mathematics, borrowing terminology to denote not-quite-analogous concepts in decision theory. This curious affectation seems also to have been the genesis of the bastardized usage of "ordinal" and "cardinal" when talking about preferences and utilities in economics).

Expected Utility Theory takes an individual's preference ordering as its input. A preference ordering is simply a list of states of the world $\langle c_0, c_1, \ldots \rangle$ such that the individual either prefers c_{i+1} over c_i or is indifferent as between c_{i+1} and c_i (this is given by the "completeness" and "transitivity" axioms.) The model further assumes that when those states of the world obtain with some probability, the individual's preferences in the resultant lotteries are stochastically consistent (this is given by the "independence" and "continuity" axioms). By varying the probability terms, the individual's indifference points may be determined (i.e., where an individual is indifferent as between some c_m and the probability p of some c_n (where $n > m$) plus the probability $(1-p)$ of some c_o where $m > o$). Using these indifference points, we can identify the values of a utility function, $u : \{c_0, c_1, \ldots\} \to [0, 1]$, assigning real values to states of the world, such that maximization of the function corresponds to the preference ordering $\langle c_0, c_1, \ldots \rangle$.

Although von Neumann and Morgenstern, 1953 is the standard formulation, it is not the only formal model of Expected Utility Theory. For example, see Alt, 1971 [1936]; Suppes and Winet, 1955. Moreover, if probabilities are meant to represent *uncertainties* rather than *risks*, Expected Utility Theory will require a subjective formulation, the standard formulation of which is *Subjective* Expected Utility Theory, axiomatized by Savage, 1954.

[12]I.e., that choosing outcome A over outcome B implies that he prefers A over B; and that choosing lottery A' over lottery B' implies that he prefers the expected outcome of A' over the

representation theorem establishes that there exists a unique (up to affine transformation) utility function, which represents his preferences for states of the world as points along the real number line.

For example, suppose there are three states of the world: A, B, and C. An individual is observed to prefer outcome A over outcome B, and to prefer outcome B over outcome C. Let us represent the utility of A with $u(A) = 1$ and the utility of C with $u(C) = 0$. The von Neumann-Morgenstern representation theorem establishes that for any real number $r \in \mathbb{R}$, there exist a probability $p \in [0, 1]$ that A will obtain, such that the expected utility $EU(p \cdot A + (1 - p) \cdot C) = p \cdot u(A) + (1 - p) \cdot u(C) = r$. And because we defined $u(A) = 1$ and $u(C) = 0$, it follows that $r = p$.

Now, because B is preferred to C, and A is preferred to B, it follows that $u(B) < EU(1 \cdot A + 0 \cdot C)$ and $u(B) > EU(0 \cdot A + 1 \cdot C)$. So there must exist a point $x \in [0, 1]$ such that $u(B) = x$. Specifically, we can represent the utility of B as the point at which the decision-maker is indifferent as between outcome B and the lottery $p \cdot A + (1 - p) \cdot C$, which will be given by $u(B) = p$.

There is much more to say about expected utility—its consistency when nested and its implicit incorporation of diminishing marginal utility among other things. These topics are thoroughly explored in a multitude of economics textbooks. For present purposes, there are two critical points. First, that ordinal preferences (i.e., that A is preferred over B, and B is preferred over C) are converted into a cardinal representation (i.e., that the utilities $u(A)$, $u(B)$, and $u(C)$ are points on a real interval). Second, that we can infer a cardinal representation of the utility of B from the expected utility of $p \cdot u(A) + (1 - p) \cdot u(C)$.

Thus, given any set of potential choices, assuming logical and probabilistic consistency, and assuming choices reveal preferences, it is possible to construct a utility function, the maximization of which corresponds to the potential behavior of any individual. More concisely: a decision-maker is "rational" iff his choices can be represented as the maximization of an expected utility function. The von Neumann-Morgenstern representation theorem establishes that such a function must exist (and tells us *how* it can be constructed) iff certain plausible assumptions are satisfied.

The von Neumann-Morgenstern representation theorem is an extraordinarily powerful result. Recasting the theorem in terms of our hypothetical "utility-maximizing robot," it is equivalent to saying that *if* a human being's behavior satisfies certain logical and probabilistic constraints, then there *must* exist some utility function, which if fed into the robot, would produce behavior identical to that human being's behavior. More than this, the von Neumann-Morgenstern representation theorem establishes that the utility function will be *unique* (i.e., that each human being's

---

expected outcome of B'.

behavior will correspond—up to affine transformation—to the maximization of *one and only one* utility function), and it suggests *how to determine* what that utility function is.

The implications which follow from the theorem are manifold. Indeed, the most interesting results in theoretical economics could fairly be characterized as merely unpacking the consequences of the von Neumann-Morgenstern representation theorem.

However, a wealth of experimental and empirical research seems to demonstrate that individuals in the real world often fail to satisfy the theorem's antecedent conditions.[13] Its "plausible assumptions" about the logical and probabilistic consistency of revealed preferences seem ever less plausible the closer we scrutinize. For example, suppose a decision-maker were presented with two decisions. In the first decision, he must choose between:

A. 50% probability of receiving \$100.

B. 100% probability of receiving \$50.

In the second decision, he is given \$100 upfront. He must choose between:

C. 50% probability of having to lose the \$100.

D. 100% probability of having to lose \$50.

Variations of this setup have been tested in very many experiments, and it has been found that most test subjects will tend to select choices (B) and (C). This is inconsistent with von Neumann-Morgenstern rationality. If individuals prefer option (B) over (A), then that means their expected utility $EU(.5\,(\$100)+.5\,(\$0)) < u(\$50)$. But if this is true, then they should choose (D) rather than (C). Experiments like this seem to reveal that individuals in the real world fail to satisfy the conditions which expected utility theory assumes.

The experimental and empirical evidence is *doubly* alarming for classical conceptions of rationality. It suggests not only that real-world behavior deviates from the premises of expected utility theory, but that it deviates *systematically*. Systematic deviation implies that expected utility theory is not merely inaccurate, but profoundly mistaken in its characterization of human decision-making. The evidence implies the existence of mechanisms not captured by the expected utility model.

It has become fashionable to name these systematic deviations with appellations like "anchoring effect," "ambiguity aversion," "availability bias," "endowment effect," "optimism bias," among a profusion of other labels.[14] The general payoff

[13] *Supra* Note 34.
[14] *Supra* Note 34.

of these various observations is not only that real-world behavior fails to converge toward rationality, but that it converges toward *something else* altogether.

To summarize the foregoing discussion: expected utility theory establishes that *if* individuals' preferences have a logically consistent structure, *and if* individuals respond consistently to probabilities, *and if* their choices reveal preferences, then their preferences may be represented as the maximization of a real-valued function. And conversely, if preferences can be represented as the maximization of a real-valued function, then they must have the structure assumed by expected utility theory. However, the experimental tests demonstrate that human behavior is systematically inconsistent, and we must therefore conclude that it *cannot* be the case that individuals' choices are representable as the maximization of an expected utility function.

Restated in terms of our hypothetical utility-maximizing robot: the experimental and empirical research suggest that for any real human being, there cannot exist *any* utility function which would cause a utility-maximizing robot to replicate the behavior of its human model.

This is a potentially devastating result.[15] If expected utility theory fails to provide a general representation of human behavior, then it leaves us with no principled basis for describing individuals' responsiveness to incentives. It renders otiose the conventional analysis of equilibria, and thereupon the whole edifice of theoretical economics crumbles.

Economics thus finds itself in the lamentable position of a science in search of a theory. Although there have been valiant attempts to rebut the experimental and empirical evidence,[16] which do much to mitigate the damage by showing that individuals *do* tend to behave rationally in *very many* situations, nevertheless the volume and validity of controverting observations ultimately seem irresistible: in at least *some* situations, people do in fact systematically behave in a manner inconsistent with the axioms of expected utility theory. At a minimum therefore this implies that expected utility theory cannot be a *complete* theory of decision-making.

2.1.2 The Menagerie of Biases and Heuristics

Given that expected utility theory is at best a partial description of human behavior, the question naturally arises whether there exists *any* general theory capable of

[15] This is sometimes presented as a problem merely for the use of cardinal utility in economic analysis. This is undersells the significance of the behavioral research. Violations of the completeness and transitivity axioms are deeply problematic for *any* conception of preference-driven behavior, regardless whether it is represented cardinally *or* ordinally.

[16] For example, Al-Najjar and Weinstein, 2009. See generally *supra* Note 4.

representing real-world decision-making?

Some scholars have answered this question skeptically. They speculate that human behavior may be too complex a phenomenon to be reducible to a tractable general model.[17]This skeptical position characterizes human behavior as an arbitrary collection of decision-making procedures (or perhaps mere brute tendencies), lacking any intelligible underlying structure: individuals exhibit the "anchoring effect" in some situations, "availability bias" in other situations, and "loss aversion" in yet other situations, etc. Some behavioralists would go further, affirmatively committing to the proposition that *no* general explanation may be constructed as to *why*—except perhaps that these cognitive traits are the byproduct of historical accidents in the course of biological evolution, no more susceptible to elucidation than why humans have ten toes rather than twelve. Although this position is rarely made explicit, the gist of it is detectable in much recent scholarship.[18]

There are many reasons to resist the characterization of human behavior as an unstructured menagerie of biases and heuristics. First, we should *want* a tractable general model of behavior if we hope to do any science. Of course, wanting a thing does not make it possible, but we should not give up the project merely because our initial attempts have failed. There is good reason to suspect—at least pre-theoretically—that a general model of behavior is obtainable. Minimal self-reflection seems to reveal that our actions *are* motivated by our preferences for their expected outcomes. It does not seem that people instinctively and unthinkingly deploy arbitrary heuristics to determine choices.[19]

Second, the menagerie conception is unlikely to withstand even basic tests. For example, one very simple test would be to ask individuals to select from two choices: (A) receiving \$10 or (B) receiving \$5. It seems utterly obvious that very nearly everyone would choose to receive \$10. I anticipate few critics would challenge this "result." Presumably, the alternatives could be varied to obtain similar results for:

2.	(A′) losing \$10	(B′) losing \$5
3.	(A″) receiving \$2	(B″) receiving \$1
4.	(A‴) receiving \$10,001	(B‴) receiving \$10,000.

[17]For example, see Tversky and Kahneman, 1992, p. 317.

[18]*Supra* Note 17. See also Jolls, Sunstein, and Thaler, 1998.

[19]Although I am inclined to regard this argument the most powerful reason to reject the "unstructured menagerie" conception of human behavior, I expect that it will be unpersuasive to many readers. It seems that very many economists have adopted a doctrinaire inflexibility about what constitutes "evidence," founded upon surprisingly superficial epistemological dogma. While it is surely the case lay people and clinical psychologists tend to overvalue introspection as a source of knowledge, the prevailing economic disdain for it is surely an overcorrection.

No matter how the payoffs are varied in this setup, it seems eminently plausible that individuals will consistently choose the dollar-maximizing alternative when choosing between certain outcomes.[20]

The only plausible deviations from maximizing behavior in these simple tests admit explanations consistent with the rational actor hypothesis. For example, an individual might choose to receive \$10 rather than \$10.01 if the inconvenience of dealing with a penny were greater than its \$0.01 value.[21] Yet bracketing off such trifling snags in experiment design, the anticipated result is so obvious that the "experiment" can hardly be thought worth conducting. *Of course* people will choose to receive more money rather than less, and to lose less money rather than more. To the extent that dollars are a tool for measuring preference magnitudes, the test implies that individuals' preferences tend to be logically consistent (i.e., that their preferences are complete and transitive). Or at least they are consistent in the pared down case of preferring more money to less money. Generalizing the claim to *all* preferences requires some work, and the implications will be fuzzy a best. Nevertheless it establishes an important insight.

If we assume that individuals value money instrumentally (i.e., as a means of achieving outcomes and not for its own sake) and if we assume that individuals are able to consistently assign monetary values to outcomes (i.e., that they are willing to pay proportionally more to achieve outcomes they prefer more), then the observed logical consistency in individuals' preferences for money will *tend* to generalize to their preferences for states of the world.[22] This implies the existence of *some* structure in human behavior and furnishes some reason to doubt the "unstructured menagerie" conception of decision-making.

Considerable care is required here. It is very easy to overstate the significance of the "more money" test. It merely *suggests* by way of a proxy that there exists some consistent principle underlying human behavior, which resembles utility-maximization. It suggests that when the elaborate decision problems posed to test subjects in experimental settings are simplified—reducing the cognitive cost of bargaining, evaluating probabilities, and grappling with unknowns—individuals do seem

[20] Indeed, the "reference point" may also be freely varied, and the result is likely to be unchanged.

[21] Other deviations—also consistent with economic rationality—may be motivated by religious, moral, or political commitments. Another possible explanation for deviation might be the suspicion that tends to accompany offers of "free money." Indeed, the enterprising psychologist might thereupon hypothesize the existence of "free money aversion," or a "too good to be true" heuristic. I would regard these as artifacts of the experiment design however, and in any case consistent with the predictions of the theory presented in this article.

[22] More precisely, the claim requires that there exist a homomorphism between non-monetary outcomes and money.

to exhibit *some* consistent behavior. They do not flap about randomly, applying arbitrary heuristics hither and thither, sometimes choosing to receive less money or to lose more. Yet all that can be inferred is that if decision problems are made sufficiently simple, then behavior will tend to exhibit *some* consistent overall structure.

Undoubtedly, many readers will find the foregoing hypothetical "experiment" trivial and uninteresting. It certainly is, but this only reinforces the point. It is *obvious* that general behavioral patterns exist—that preferences for outcomes *do* tend to drive decision-making.

And of course no experimentalist possessing a scintilla of ambition or creativity would bother to conduct the "more money" test, precisely *because* its outcome is trivial. For this reason, there exist few publications "confirming" rational behavior in such reductive setups.[23] Meanwhile, scholarly journals are replete with studies revealing the ineptitude of test subjects facing strange and unnatural decision problems. Given the incentives to produce novel scholarship, behavioral research will tend to generate an exaggerated impression of how badly the received view fails. Indeed, research in *any* scientific field will tend to generate an exaggerated impression of the defects of *any* received view. It is entirely understandable why so many economists would feel inclined to adopt a nihilistic attitude toward general theories of behavior, given the volume of seemingly unrelated, inconsistent, and self-defeating cognitive biases uncovered in the lab. The sheer abundance of experimental articles reporting observations of non-rational behavior are liable to create the impression that human behavior is considerably less ordered than it is in fact.

But let us not lose perspective. The rational actor hypothesis was not conjured from thin air, but induced from robust patterns observed informally in everyday life. Meanwhile, the observations gathered from lab experiments are invariably elicited by hurling probability riddles upon test subjects. Typically, these test subjects are hapless undergraduates, the selection of whom deliberately excludes those majoring in economics, statistics, or any other art which might prepare them to grapple with such problems. And inasmuch as money-seeking behavior does not straightforwardly imply rationality, it is a far weaker inference from the existence of observed anomalies to the conclusion that human behavior has *no* general structure.

Now, lest I be misunderstood, I do *not* mean to imply that experimentalists ought to devote more attention to reductive experiments like the "more money" test, nor indeed that they should design experiments with the aim of "verifying" the axioms of expected utility theory. Clearly, the *interesting* experiments are those which yield controverting observations. The point is simply to be mindful that the *reason why* observations of non-rational behavior are scientifically interesting

[23] Some examples are discussed in Zamir, 2018.

is precisely *because* strong evidence of rational behavior would arise in reductive experiments. An economist disaffirming the rational actor hypothesis on the basis of the anomalies thus far observed would be like a physicist doubting the law of gravitation after observing examples of buoyancy, aerodynamic lift, and magnetism. This is not to say that the persistent observation of anomalies cannot undermine our belief in a proposed general law, but merely that we should exercise caution when jettisoning a theory that seems to work in *most* instances—especially in the absence of an acceptable replacement.

Another reductive test of the menagerie conception would be to provide test subjects with an "expected utility calculator" in any of the experiments where non-rational behaviors have previously been observed. Suppose the calculator accepts two inputs: the utility that test subjects assign to states of the world and the probabilities associated with each outcome attached to a given choice. Using these inputs, the calculator outputs the test subjects' utility-maximizing choices. The relevant question is whether, given the results produced by the calculator, test subjects would continue to exhibit logically or probabilistically inconsistent behavior. Would test subjects select the utility-maximizing choice if they were *told* which alternative were utility-maximizing?

Again, assuming away artifacts of experimental design—i.e., assuming test subjects understand what the calculator does and how to use it—it seems *obvious* that access to the calculator would dramatically reduce observations of non-rational decision-making. Of course test subjects will select the utility-maximizing choice if they are told which choice is utility-maximizing. The implications of this are nontrivial. The "expected utility calculator" test implies the *source* of non-rational decision-making. It reveals that individuals *do want to* maximize, but simply lack the cognitive tools required to do so effectively in scenarios where they lack access to an "expected utility calculator." Indeed, this intuition seems to be precisely what motivates experimentalists to exclude economics students from their testing pools. Trained individuals are less susceptible to the inconsistent behaviors which experimentalists seek to uncover. But this concedes that people generally would employ the tools of economics and statistics to make optimal choices if only they had access to that knowledge.

The "more money" test and "expected utility calculator" test remind us of our theoretical baseline and help us to identify what is marginal (and why it is marginal). The extravagant conclusion which skeptics draw from the experimental and empirical literature—that individuals do not systematically seek outcomes which maximally satisfy their preferences—is not justified on the ground of such margin-pushing experiments.

I have perhaps devoted more discussion to rebutting the menagerie model (or *anti*-model as the case may be) than some might think warranted. It seems that a majority of theorists still believe that a successful general theory may yet be devised. Nevertheless, the pervasiveness of the nihilistic view ought not be underestimated, and insofar as *some* readers may hold such a view, the foregoing remarks are worth stating expressly.

2.1.3 Remedial Theories

If a general model of behavior *can* be devised, then the next question is what such a theory might be. The predominant approaches toward a general descriptive theory of decision-making can be divided into two intersecting classes: (i) theories of bounded rationality, and (ii) generalized expected utility theories.

The relationship between theories of bounded rationality and generalized expected utility is rarely made explicit. This has led to some confusion. Exploring their relationship is useful to understanding how behavioral decision theory has developed. The two categories are neither coextensive nor disjoint. Some (but not necessarily all) generalized expected utility theories are theories of bounded rationality, and some (but not necessarily all) theories of bounded rationality are generalized expected utility theories.

Theories of bounded rationality provide an intuitive explanation *why* real world decision-making often appears to diverge from the predictions of rational choice theory, whereas generalized expected utility theories emend expected utility theory to construct models which more closely represent real world behavior. Theories of bounded rationality might be characterized as offering an answer to the apparent failure of the strong rational actor hypothesis, and generalized expected utility theories as offering an answer to the apparent failure of the weak rational actor hypothesis.

This requires further elaboration, yet before embarking upon a thorough treatment of that relationship, we should first get clear on what precisely "bounded rationality" and "generalized expected utility theories" denote.

Bounded Rationality

All theories of bounded rationality begin from the central premise that decision-making is costly.[24] These costs include search, cognitive strain, computation, and the opportunity cost of deliberation, among other things. I will refer to these costs generally as "decision costs."

[24]The concept of bounded rationality was first described in Simon, 1955.

Explicitly defined, a "decision cost" is any cost not arising from the outcome of a choice, but rather arising from the act itself of *choosing*. Decision costs are a function of two variables: the *set of facts* which give rise to the decision, and the *method* that the decision-maker uses to select a choice.

For example, one *possible* method of selecting a choice is "rational deliberation": the decision-maker could survey the expected utility of each alternative and select the action which returns the greatest value. This of course is the method described by expected utility theory. However, rational deliberation is not the *only* method available to a decision-maker. He could choose the alternative with which he is most familiar, or which minimizes risk, or which minimizes the disutility of the worst case, or which maximizes the utility of the best case. The possible ways of choosing are practically limitless.

Let "decision-making procedure" denote any possible method that a decision-maker can conceivably employ in selecting a choice. More formally, a decision-making procedure is a mapping from states of the world to choices. The set of decision-making procedures includes rational deliberation, heuristics, coin flips, and any other conceivable method used to determine a choice.[25]

Theories of bounded rationality are plausible, because the presence of decision costs can explain why rational decision-makers *seem* to behave non-rationally. Rational deliberation tends to require comparatively greater investment in decision-making. Individuals may therefore be better off accepting a reduction in the expected utility of outcomes in order to save on decision costs. In other words, it is sometimes rational to choose not to be "rational."[26]

To illustrate the point, consider the following decision problem: Select from among the three alternatives below. Take it as given that each choice refers to a distinct monetary value not equal to any other alternative, but that no alternative varies from any other by more than two cents.

(A) Receive $\lfloor 10^{12235} \cdot \pi \rfloor - 1000 \lfloor 10^{12232} \cdot \pi \rfloor$ cents.

(B) Receive x cents, where x is the largest twin prime such that $x \leq 2023$.

[25] Rational deliberation and heuristics do not exhaust the set of all decision-making procedures. The literature typically defines "heuristics" as being decision-making procedures which tend to select worse choices than rational deliberation (i.e., at least some members in the set of potential choices will yield lower payoffs than the potential choices selected by rational deliberation), but which also incur less decision cost than rational deliberation. However, we can of course *conceive* of methods of decision-making which select worse choices than rational deliberation *and* incur *more* decision cost.

[26] This is approximately what Simon had in mind with the concept of "satisficing." Simon, 1956.

(C) Receive y dollars, where y is the number of times the word "their" appears in the United States Declaration of Independence.

Most individuals faced with this problem would presumably select an arbitrary choice, because a possible two cent improvement in payoff cannot justify the tedious work of determining the optimal choice. It is not worth the effort to work out which of the three alternatives denotes the greatest monetary value.[27]

Implicit in theories of bounded rationality is the recognition that the selection of a decision-making procedure is itself a decision. Individuals do not simply select actions, they must also choose *how to choose* a possible action. Choosing *how to choose* is itself a decision. Let us call the choice of action a "first order" decision, and the choice of first order decision-making procedure a "second order" decision.[28]

Arranging decisions hierarchically in this way, it becomes sensible why rational actors might choose not to be "rational." It can be second order rational not to choose to be first order rational. Given the prohibitive decision costs associated with optimization, *first order* rational deliberation might perform worse than some other first order decision-making procedure on balance.

Thus does bounded rationality explain apparent deviations from the predictions of expected utility theory. And it does so while preserving the kernel of the strong rational actor hypothesis: that decision-makers *are* optimizing. They are optimizing over decision-making procedures rather than actions. Yet their behavior is still fundamentally *optimizing*. They are still essentially *rational*. If a rigorous second order theory of rationality were constructed, then we might reasonably suspect that it could explain the "anomalies" observed in the experimental and empirical literature, while preserving the basic conceptual framework of utility-maximization.

Yet two profound problems are immediately apparent. First, if we are to understand every decision as analyzable into two distinct decisions—i.e., (i) what to do, and (ii) how to decide what to do—then this begs the question: is the decision *how to decide* how to decide not also a decision? And is the decision *how to decide* how to decide how to decide not another decision still?

The bounded rationality approach kicks the lid off Pandora's box, for if decision-making procedures are themselves taken to be choices, then every first order decision

[27]The payoffs are (A) \$20.01, (B) \$19.99, and (C) \$20.00. I assume of course that the decision-maker does not derive any pleasure from working out the values, although if he did, then this might be characterized as a *negative* decision cost (or maybe a "decision utility"). Negative decision costs are not addressed any further in this article, but they are compatible with the model I present.

[28]Few authors have treated the layering of higher order decisions explicitly. Sunstein and Ullmann-Margalit, 1999 describe something approximately equivalent to my conception.

implies not only a second order decision, but rather an *infinitude* of higher order decisions.

Decision	Choice Set
...	...
n^{th} order	Methods of selecting an $(n-1)^{\text{th}}$ order choice.
...	...
Third order	Methods of selecting a second order choice.
Second order	Methods of selecting a first order choice.
First order	Actions.

The abyss of an infinitude will naturally tend to elicit some discomfort, however this is not by itself fatal. What *does* render the problem dire is the premise, essential to theories of bounded rationality, that all decision-making is *costly*. There does not seem to be any principled reason to suppose that the decision costs associated with each i^{th} order choice converge to zero as $i \to \infty$. We are left therefore with the intolerable result that *all* decisions presumptively imply *infinite* decision costs. This is not progress, for we have moved from trying to explain apparently irrational decision-making to the absurd conclusion that decision-making is impossible. That cannot be right.

Surprisingly few theorists of bounded rationality have attempted to address this problem.[29] To avoid the issue, they simply *assume* a finite hierarchy of higher order decision-making. More specifically, nearly all models of bounded rationality assume exactly *two* levels of decision-making, surmising decision-makers to be second order rational. No justification is offered for this restriction.

It is difficult to fathom how bounded rationality could be regarded a satisfactory general theory of behavior without addressing the infinite regress problem. The core conceptual move cannot avoid question-begging in the absence of an explanation why the model should be delimited to a finite hierarchy, or why we should believe decision costs to be convergent when iterated to infinity. Thus arises the first major obstacle to developing a theory of bounded rationality.

Yet even putting aside the infinite regress problem, there exists a *second* major obstacle. Bounded rationality hypothesizes that for any i^{th} order decision, where $i > 1$, decision-makers are faced with the choice: which $(i-1)^{\text{th}}$ order decision-making procedure to employ. If the decision, for any i, depends on the expected utility of each $(i-1)^{\text{th}}$ decision-making procedure less its decision costs, then the

[29] Despite the dearth of attempts to resolve it, behavioral economists are certainly acutely *aware* of the problem. *See* Conlisk, 1996. *See also* Minsky, 1986, referring to the problem as "Fredkin's paradox."

decision-maker would need to *determine* what the expected utility of selecting that decision-making procedure would be. But if the decision-maker were to run through the deliberation necessary to determine the expected utility of using every available $(i-1)^{\text{th}}$ order decision-making procedure, then he would incur the very decision costs (and more) which higher order rationality was posited to avoid. Specifically, he would incur the decision cost of determining the expected utility of every alternative *plus* the decision cost determining the decision costs of every alternative *plus* the decision cost of selecting a maximum.

This is ugly but not yet calamitous. Choice sets (of any order) can be constrained to finite cardinalities. Plausible arguments may be raised for adopting a finiteness assumption, and this avoids the threat of yet another infinite regress. But what is not so easily fixed is far more profound: that it is never $(i+1)^{\text{th}}$ order rational to choose i^{th} order rationality, for any $i \in [2, \infty)$.

The decision cost of rational deliberation for any i^{th} order of decision-making, where $i > 1$, must be greater than the decision cost of $(i-1)^{\text{th}}$ order rational deliberation. Therefore, i^{th} order rational deliberation cannot have expected payoffs greater than $(i-1)^{\text{th}}$ order rational deliberation. But if i^{th} order rationality returns worse payoffs than $(i-1)^{\text{th}}$ order rationality, then it would be $(i+1)^{\text{th}}$ suboptimal to choose to employ i^{th} order rational deliberation. It would be suboptimal, because there will always be at least one i^{th} order decision-making procedure which returns a better payoff: i.e., simply to *do* the $(i-1)^{\text{th}}$ order rational deliberation without weighing whether it would be optimal.

It is important to recognize here that "choosing $(i-1)^{\text{th}}$ order rational deliberation" is not identical to i^{th} order rational deliberation. It could be i^{th} order suboptimal to choose $(i-1)^{\text{th}}$ order rational deliberation. The i^{th} order choice to be $(i-1)^{\text{th}}$ order rational can be $(i+1)^{\text{th}}$ order rational, even if it is not i^{th} order rational.

The problem then, stated more succinctly, is that any i^{th} order rational deliberation will subsume $(i-1)^{\text{th}}$ order rational deliberation. Deliberating whether it would be i^{th} order rational to be $(i-1)^{\text{th}}$ order rational requires undertaking the $(i-1)^{\text{th}}$ order rational deliberation, and therefore it will always be inferior to *just doing* the $(i-1)^{\text{th}}$ order rational deliberation without considering alternatives. In other words, higher order rationality is redundant—it necessarily duplicates all the decision costs of lower order rationality. Therefore, it can only *add* decision costs by considering non-rational $(i-1)^{\text{th}}$ order alternatives, undermining its very *raison d'être*.

And of course, the experimental and empirical evidence has established that at the ground level, individuals *do not* use first order rational deliberation. Therefore, no individual is i^{th} order rational at *any* order of decision-making i, contradicting the

principal claim of bounded rationality: that individuals are rational at some order of decision-making.

Direct description of the problem is admittedly somewhat abstract, and a numerical example may help ground intuitions. Suppose an individual were planning to travel from New York to Washington, and he had to choose whether to journey by train or by air. He has several possible second order choices. He could weigh the various factors: speed, convenience, comfort, price, risk of accidents, risk of delay, etc., to determine the expected utility-maximizing choice. That would be a first order rational deliberation. Alternatively, he could choose whichever mode of transit sprung to mind first. Or he could choose whichever mode of transit were cheaper. Or he could choose whichever mode of transit were faster. Or he could choose the closer local destination: the train station or the airport. Or he could flip a coin.

How should the decision-maker select from among these alternatives? If the individual were second order rational, then he would calculate the net expected utility of each decision-making procedure, subtract their decision costs, and choose the optimal decision-making procedure. He would then apply the optimal decision-making procedure to determine an action.

Suppose the expected utility of traveling by airplane were 10 and the expected utility of traveling by train were 5. And suppose further the following values:

Decision-Making Procedure	**First Order Expected Utility**	**Decision Cost**	**Second Order Expected Utility**
Rational Deliberation	10	8	2
Cognitively Salient	5	1	4
Cheapest Alternative	5	3	2
Fastest Alternative	10	7	3
Coin Flip	7.5	1	6.5

To get a sense of the mechanics of second order rationality, let us consider the third alternative: "choose the cheapest alternative." The expected utility of choosing the cheapest alternative is 5 (given in the second column), and the cost of *determining* the cheapest alternative—i.e., the burden of comparing prices and hunting for discounts—is 3 (given in the third column). Therefore, the second order expected utility of choosing the "choose the cheapest alternative" decision-making procedure is $2 = 5 - 3$ (given in the fourth column).

This alternative is plainly not the second order expected utility-maximizing procedure, since there are other decision-making procedures which return greater second order expected utility values. The second order *rational* decision-maker would choose

the decision-making procedure which returns the greatest second order expected utility value. That would be the "coin flip" decision-making procedure, which returns a second order expected utility of 6.5.

But now consider: in order to *ascertain* that the second order expected utility of using "coin flip" is 6.5, the decision-maker would have to *determine* the first order expected utility and decision cost of the coin flip. And in order to determine that it is the optimal decision-making procedure, he would have to determine the second order expected utilities of all the alternative decision-making procedures. And in order to do this, he would have to determine all of their first order expected utilities and decision costs. And these are the same decision costs (and more) that he would have had to undertake if he were simply *first* order rational.

Thus, the *second order* decision cost of rationality would be the sum of *all* the first order decision costs (in the example, $8 + 1 + 3 + 7 + 1 = 20$). The second order rational individual would therefore experience a (third order) expected utility of $-12.5 = 7.5 - 20$.[30] Yet this leads to an absurd result. If the decision-maker were *simply* first order rational, then he would experience an expected utility of 2, which is greater than the utility of employing second order rational deliberation. After all, $2 > -12.5$. Rather than investing effort to determine the best decision-making procedure, individuals would be better off simply engaging in first order rational deliberation spontaneously. Certainly, "do a first order rational deliberation" is an available second order decision-making procedure, and assuming it incurs negligible decision costs, it will always be superior to second order rational deliberation. Indeed, under these assumptions it is trivially the case that first order rational deliberation will *always* return greater utility in the aggregate relative to any i^{th} order rational deliberation, for $i > 1$. Thus, we are left where we started: a theory of first order rationality, which the experimental and empirical evidence has already controverted.[31]

More problematically still, the foregoing considerations generate a *doubly* intolerable result. If the decision-maker has incurred the decision costs necessary to determine the payoffs of employing every available decision-making procedure, then although he may conclude that it *would have been* second order optimal to choose

[30] Of course, some of the decision costs of the alternatives may be overlapping. It may be the case, for example, that the decision-maker can determine the cheapest alternative or fastest alternative at less cost *after* having already undertaken rational deliberation. Nevertheless, it would still be the case that the second order decision-making procedure would incur *at least* the cost of first order rational deliberation and whatever second order deliberation is required to compare alternatives.

[31] The problem I have just described has some antecedents in related fields. In macroeconomics, there is some relationship with "rational expectations," and in operations research, it bears some similarity to the "decision-making paradox" described in Triantaphyllou and Mann, 1989.

some suboptimal first order procedure in order to save on decision costs, it would be preposterous to actually adopt that decision-making procedure in selecting a choice. When he has formed this determination, he has *already* incurred the decision cost of first order rational deliberation. He had to perform the first order rational deliberation in order to know its payoffs and to see that it was not second-order optimal. But having done this, he knows the optimal action. Knowing the optimal action, it would be ridiculous to then choose a suboptimal action, on the specious ground that the decision-making procedure which determined the suboptimal act would be second order optimizing. The coin flip is second order optimal because it saves on decision costs, but those decision costs have already been "spent" determining this information.

More concisely: in order for decision-makers to decide second order rationally which first order decision-making procedure to use, they must weigh the payoffs of using each decision-making procedure against its decision costs. However, in order to gauge the payoffs of a decision-making procedure, decision-makers will need to determine the utility value of its output. And determining the utility value of its output is the very thing that notions of higher order rationality were posited to avoid. Therefore, second order optimization cannot improve on first order optimization. And this generalizes to higher order rationality straightforwardly. For convenience, let us refer to this problem as the "idempotent hierarchy problem."

Generalized Expected Utility Theories

Whereas theories of bounded rationality propose substantive claims about the psychology of decision-making, generalized expected utility theories tend to focus instead on predictive validity. Broadly, generalized expected utility theories attempt to preserve the maximization hypothesis of classical rationality while altering the formulation so as to better conform with experimental observations. The fundamental premise which these theories share is a plausible one: that the utility functions described by the von Neumann-Morgenstern theorem are too simplistic to represent how real people evaluate states of the world. Yet this does not necessarily entail that human behavior could not be represented as the maximization of some *different* function.

The defining objective of generalized expected utility theories is to supply a more empirically grounded formulation of utility maximization. The intuition is that decision-makers may be represented as maximizing *something*. Whether this position is tenable depends upon *what* that "something" is, and whether the alternative formulation proves to be more consistent with observed behavior than expected utility

theory.

Experimental work has tended to focus on test subjects' inconsistent choices under conditions involving risk and uncertainty. Accordingly, generalized expected utility theories typically focus on reformulating the probability functions of expected utility theory, informed by the results of experimental observations.

The most influential generalized expected utility theory is "prospect theory," first exposited in Kahneman and Tversky, 1979.[32] Prospect theory introduces three complications to the expected utility model. First, utility is characterized not as a function of states of the world, but rather in terms of gains and losses. Test subjects have been observed to be more than twice as sensitive to perceived losses than to perceived gains. Therefore, prospect theory posits that *changes* to the state of the world with respect to a reference point determines a subject's "value function."

Second, probabilities are subject to underweighting and overweighting when contributing to a choice. Test subjects have been observed to behave *as if* a probability were greater or less than its given value. These variations were not due to misestimations or miscalculations, but rather an apparent quirk in the operationalization of known probabilities. Prospect theory thus posits a weighting function, which modifies perceived probabilities.

Together, the value function, v, and probability weighting function, π, describe "prospects," which are sets of ordered pairs $\{\langle v(c_0), \pi(p_0)\rangle, \langle v(c_1), \pi(p_1)\rangle, \ldots\}$, where c_i is a state of the world and p_0 is the probability of its obtaining. The value function describes a discontinuous 'S'-shaped utility curve, and the probability weighting function overweights lower probabilities and underweights higher probabilities. The prospects thusly formulated are then "edited" with various heuristics to simplify the decision problem before decision-makers maximize their payoff from the modified "expected utility function."

The foregoing description admittedly glosses over many nuances. And there have emerged several competing generalized expected utility theories which resist analogy to prospect theory. Nevertheless, the rough and ready summary suffices for the purposes of illustration.

In their favor, generalized expected utility theories have the benefit of being "empirically grounded" *ab initio*. Like prospect theory, nearly all generalized expected utility theories start from a base of experimental observation. However, they suffer from two critical drawbacks.

First, they fail to provide much explanatory insight as to *why* individuals seem so inept at evaluating outcomes and reasoning probabilistically. Prospect theory posits that individuals behave as if they were maximizing over sums of values discounted

[32] See also Tversky and Kahneman, 1992.

by probability weighting functions, but it supplies no insight as to what might cause such behavior. It simply *describes* how test subjects were observed to behave. It does not *explain*. This is a repairable defect of course, for if an otherwise satisfactory generalized expected utility theory were discovered, then it might reasonably be anticipated that—with some ingenuity—explanatory intuitions might be extracted from it.

The merit of a generalized expected utility theory thus depends almost wholly upon its capacity to predict behavior. Yet this leads to the second complaint: that when the parameters of the very experiments upon which generalized expected utility theories are based are varied, their models seem to be no more resilient than classical expected utility theory at predicting behavior.[33]

The second defect is dispositive, but the first is still worth further discussion. One approach to mitigating the explanatory problem, which some generalized expected utility theorists have proposed, involves embedding generalized expected utility theories within the broader conceptual framework of bounded rationality. Given that probabilistic reasoning is cognitively onerous, it is plausible that when facing situations involving risk or uncertainty, decision-makers will more readily avail themselves of non-rational (and accordingly less costly) decision-making procedures. The probability weighting function of prospect theory, for example, might represent an aspect of some commonly employed decision-making procedure. This approach has several advantages and several disadvantages.

In its favor, embedding a generalized expected utility model within a conceptual story of bounded rationality can provide a plausible intuitive foundation. By characterizing a generalized expected utility theory as modeling a particular decision-making procedure which individuals deploy when facing risk or uncertainty, there is a compelling story *why* decision-makers might exhibit probabilistically inconsistent behavior. Optimizing under conditions of risk and uncertainty entails high decision costs. Consequently, the higher order rational decision-maker might plausibly choose to adopt the decision-making procedure described by a generalized expected utility theory when facing those risky or uncertain decisions.

Yet embedding generalized expected utility theories within the explanatory framework of bounded rationality has drawbacks. Recall that theories of bounded rational-

[33]For example, Birnbaum has published a number of experimental papers testing prospect theory and its variants, offering compelling evidence of its fragility. See, e.g., Birnbaum, 2004, 2006, 2008a, 2008b; Birnbaum and Navarrete, 1998. Birnbaum's results seem to suggest that alternative generalized expected utility theories are considerably more durable, however these too have fared poorly when tested against novel experimental designs. Researchers advancing this line of inquiry may view these rejected models not as failed enterprises, but rather as hopeful steps toward a general theory approached incrementally. I am doubtful of the experiment-driven theory approach.

ity encounter obstacles of their own: the infinite regress problem and the idempotent hierarchy problem. The use of bounded rationality as an explanatory support for generalized expected utility theories is thus a double-edged sword.

Moreover, simply "embedding" the two approaches without rigorously formulating a complete general framework leaves too much unanswered. Why should anyone believe a proposed generalized expected utility theory describes how individuals reason about *all* decisions? Why would decision-makers choose the procedure described by prospect theory rather than rank-dependent expected utility theory, or any other deliberative mechanism? Why not mix and match a variety of decision-making procedures? If decision-makers do mix and match, then what other decision-making procedures do decision-makers use? When do they choose them? And why?

In the absence of a rigorous and complete formulation of the broader bounded rationality framework, the proposal of a generalized expected utility theory amounts to a bald assertion that such-and-such decision-making procedure is always second order optimal and that individuals are second order rational. Merely "embedding" without more is an admission that the generalized expected utility theory is not fully general. Moreover, if a generalized expected utility theory purports to be a complete theory of behavior, then it begs the question why other decision-making procedures are necessarily worse. And if it is conceded to be an incomplete theory of behavior, then it leaves unanswered the most pressing question—what conditions trigger deployment of the decision-making procedure described by that theory?

Of course, the generalized expected utility theorist could sidestep all this. Generalized expected utility theories do not *need* to provide any psychological account of decision-making. The generalized expected utility theorist may simply concede the criticism—that his theory has no intuitive basis—but insist nevertheless that it is good at prediction, and that prediction is all that matters.

Regardless, although the explanatory weakness of generalized expected utility theories is interesting, the point is moot. The generalized expected utility theory approach relies fundamentally upon a theory's capacity for prediction, and no theory yet advanced succeeds in generating good predictions consistently.

* * *

To harden intuitions, it is worth restating the goals of bounded rationality and generalized expected utility theories in terms of our hypothetical utility-maximizing robots. Both may be understood as attempts to salvage the claim that robot doppelgängers—whose decisions would be identical to those of a human—*can* be constructed.

The principal claim of bounded rationality is that a utility-maximizing robot's behavior could be made indistinguishable from a human's if it were allowed *two* initializing inputs: (i) a utility function, *and* (ii) a decision cost assignment (assigning disutility values to decision-making procedures).

The principal claim of generalized expected utility theories is that a robot's behavior could be made indistinguishable from a human's if its *treatment* of the utility function were something distinct from the straightforward maximization of expected utility.

2.2 A Meta-Model of Meta-Rational Choice

The decision-making structure presented in this section is both a theory of bounded rationality and of generalized expected utility. Importantly, it remedies all four of the major defects identified in the previous section. *Qua* theory of bounded rationality, it provides an endogenous answer to the infinite regress problem and idempotent hierarchy problem. *Qua* generalized expected utility theory, it embeds the deviations from expected utility theory within the framework of bounded rationality *explicitly*. It requires no *ad hoc* manipulation of the probability function, and it is necessarily consistent with observed behavior.

The theory is expressed in the form of a *meta*-model. It is a framework for generating representations of decision-making behavior—a model *describing models* of decision-making. Formally, a "model" in the meta-model is an ordered triple, containing (i) a set of all possible actions, (ii) a utility function, and (iii) a decision cost assignment.

The main result is that for any set of potential behaviors, there exists a model representing those chosen behaviors as optimal. The intuition is that decision-makers' behaviors can be represented by some decision-making procedure capable of generating choices for every possible state of the world, which is i^{th} order rational. However, the value of i is not assumed arbitrarily in my meta-model. Rather, the value of i is itself taken as the result of optimization.

My strategy is to construct a set containing every possible decision-making procedure of any hierarchical order. Each of these decision-making procedures will have some representative utility value (it's "meta-utility"). I prove that for any set of potential actions, there must exist a model and a meta-utility-maximizing decision-making procedure within that model, which selects those actions.

The structure of my exposition is straightforward. First, I define decision-making and states of the world (§2.2.2). Second, I define a hierarchy to represent higher order decisions (§2.2.3). Third, I define "meta-utility" and "meta-rationality" in formal

terms (§2.2.4), proving that any set of behaviors may be characterized as the product of a meta-rational procedure in some model.

2.2.1 Notation

Consistent with convention, functions are treated as sets of ordered pairs. Functions may be defined either explicitly or in relational terms, depending on ease and clarity of notation. For example, the function mapping integers to their squares $f : \mathbb{Z} \to \mathbb{Z}$ may be defined either by $f(x) = x^2$ or equivalently by $f = \{\langle x, y\rangle \mid y = x^2\}$. The variable term f is used to denote functions. If two variables are needed to denote functions, prime notation is introduced. For example, f, f', f'', etc., denote distinct functions. The terms g and h name particular functions defined below, and they will never be used as variable designations.

In general, the terms x, y, and z are used to signify variable objects—with subscripts adopted in case more than three variables are needed. The variable term S will be used to designate sets—again with subscripts adopted in case more than one variable is needed. The terms i, j, k, and l designate index variables. If a symbol is intended to designate a universal variable for a specific kind of object which ordinarily has a subscripted name, then the term may be used without subscripts to indicate that it designates any object of that type. For example, where $a_0, a_1, \ldots \in A$ name specific members of A, it may be understood that the terms a or a_i in the formulas $\forall a \in A\,(a = y)$ and $\forall a_i \in A\,(a_i = y)$ are, respectively, equivalent to writing $\forall x \in A\,(x = y)$. For objects with subscripted names, the subscript n or m will be used to designate particular variables. For example, $\exists a_n\,(a_n = x)$.

Note further that the use of superscripts should be read as an alternative to subscripts as a means of recycling base terms. Superscripts will only denote exponentiation when the base is a number. For example, x^2 simply denotes some object distinct from x^0 and x^1—it does not denote $x \times x$. However, 2^x *does* denote exponentiation.

As a general rule, subscripts, superscripts, overbars, and prime notation do not designate operations. They are simply used to name terms and variables. The only exception is when the base term is a number, in which case superscripts should be read as exponentiation.

Importantly, a common structure which arises in the meta-model is the ordered pair $\langle f, i\rangle$. For convenience, the notation f_i is introduced to abbreviate $\langle f, i\rangle$. It is especially important in this context to recognize that f_i and f_j do not denote distinct functions. Rather, they denote ordered pairs, the first element of both being the same function. More precisely, they abbreviate $\langle f, i\rangle$ and $\langle f, j\rangle$, where the first element

in both ordered pairs is the identical function f. Two objects of that form, which are distinct in both their elements, would be denoted f_i and f'_j, where f_i abbreviates $\langle f, i\rangle$ and f'_j abbreviates $\langle f', j\rangle$.

Also, as a general guide to symbol choice, the general convention I have adopted is to signify increasing generality with increasingly elaborate notation. For example, the meta-model defines $a_0, a_1, \ldots \in A$, and $A_0, A_1, \ldots \in \mathfrak{A}$.

The notation is meant to emphasize clarity over rigor. Thus, for example, when defining $a_0, a_1, \ldots \in A$ and $A_0, A_1, \ldots \in \mathfrak{A}$, the question *which* $A \in \mathfrak{A}$ contains $a_0, a_1, \ldots$, is meant to be understood from context. The obvious alternative would have been to disambiguate by defining $a^0_0, a^0_1, \ldots \in A_0$ and $a^1_0, a^1_1, \ldots \in A_1$, and $A_0, A_1, \ldots \in \mathfrak{A}$. However, this would add unnecessary notational clutter, as only one $A \in \mathfrak{A}$ is ordinarily relevant. Common sense and common conventions are otherwise assumed throughout.

2.2.2 Actions and States of the World

Let $c_0, c_1, \ldots$ represent "conditions" (or "states of the world"). Conditions are the simplest elements of the meta-model. Let $\mathfrak{C} = \{c_0, c_1, \ldots\} \neq \varnothing$ represent the set of all possible conditions.

Let "actions" be represented as ordered pairs, consisting of "antecedent conditions" and "consequent conditions," with the form $\langle c, \{\langle r, c'\rangle, \langle r', c''\rangle, \ldots\}\rangle$, where $c, c', c'', \ldots \in \mathfrak{C}$ and $r, r', \ldots \in \mathbb{R}$, such that $r + r' + \ldots = 1$.

The intuitive interpretation is that an action is a lottery which moves individuals from some antecedent condition to consequent conditions with some probabilities. In other words, $\ulcorner\langle c, \{\langle r, c'\rangle, \langle r', c''\rangle, \ldots\}\rangle\urcorner$ represents an action performed when the state of the world is c, which leads to the outcome c' with probability r, to outcome c'' with probability r', and so forth.

Definition 2.2.1 (Alternatives)**.** *For any sets x_0, x_1, y_0, y_1, z_0, and z_1, Let $z_0 \sim z_1$ iff $z_0 = \langle x_0, y_0\rangle$, $z_1 = \langle x_1, y_1\rangle$, and $x_0 = x_1$.*

The foregoing definition implies that if two possible choices, $\langle c_i, S_0\rangle$ and $\langle c_j, S_1\rangle$, share the same antecedent condition, i.e., that if $c_i = c_j$, then they are "alternatives" if $S_0 \neq S_1$ (obviously, they are simply identical if $S_0 = S_1$). The "choice set," given some state of the world, c_n, is the equivalence class formed by the $\sim$ relation on the set of all possible choices, such that the first member of every element of the equivalence class is c_n.

Definition 2.2.2 (Choice Functions)**.** *For any set S, let $G(S) = S \to \bigcup S$ denote the function space of choice functions $g_0, g_1, \ldots \in G(S)$ on any set S, for which $\forall x \in S\,(x \neq \varnothing)$.*

Definition 2.2.3 (Set of Representations)**.** *Given any set of possible conditions $\mathfrak{C}$, let the set of all ordered pairs that could represent actions be defined by:*

$$\overline{\mathfrak{A}}(\mathfrak{C}) = \mathfrak{C} \times \left\{ x \mid x \subseteq \mathbb{R} \times \mathfrak{C} \wedge \sum_{\langle y,z\rangle = i \in x} y = 1 \right\}$$

Now, let $\mathfrak{A}(\mathfrak{C})$ denote the set of all valid *representations of possible actions:*

$$\mathfrak{A}(\mathfrak{C}) = \left\{ x \;\; \mid \exists g \left(x \subseteq \overline{\mathfrak{A}}(\mathfrak{C}) \wedge g \in G(\overline{\mathfrak{A}}(\mathfrak{C})/\sim) \wedge \mathit{Img}(g) \subseteq x \right) \right\}$$

The objective here is to define a set containing all valid representations of actions. For any state of the world $c \in \mathfrak{C}$, there should exist at least one action a_i such that the actor can choose a_i if c obtains (i.e., there exists some S, such that $a_i = \{c_j, S\}$ for any $c_j \in \mathfrak{C}$).[34] Of course, there may be *more* than one possible action available in any given antecedent condition—but there must be *at least* one possible action if a set $A_n \in \mathfrak{A}(\mathfrak{C})$ is meant to represent the set of all possible actions. The intuition here is that no state of the world exists where an individual inhabiting that state of the world has *no* possible choices.[35] This is because in any state of the world, he must have at least one possible course of action (although when there is only one choice, this admittedly does some abuse—albeit harmless for present purposes—to the words "choice" and "decision.").

Now, while some A_n may exhaust the set of possible actions for a particular decision-maker at a particular time, there are of course other conceivable representations (i.e., other sets of possible actions, $A_k, A_l, \ldots \in \mathfrak{A}(\mathfrak{C})$). Definition 2.2.3 defines the set $\mathfrak{A}(\mathfrak{C})$, and we may be assured that there exists some $A_n \in \mathfrak{A}(\mathfrak{C})$, such that A_n represents the set of all possible actions of any individual decision-maker in any possible state of the world.

As a matter of notational convention, such representations will be denoted by uppercase $A_0, A_1, \ldots \in \mathfrak{A}(\mathfrak{C})$. Particular actions will be denoted by lowercase $a_0, a_1, \ldots \in A \in \mathfrak{A}(\mathfrak{C})$.

[34] The condition given by the expression $\exists g \left(g \in G(\overline{\mathfrak{A}}/\sim) \wedge \mathit{Img}(g) \subseteq x \right)$ is meant to ensure that for any state of the world, there exists at least one choice alternative available to the decision-maker.

[35] At the margins, it seems that even where it may *seem* that an individual cannot *act*, it must at least be the case that "doing nothing" is a possible action—perhaps in some cases the only possible action. Even in such cases, there exists a possible choice (i.e., to do nothing).

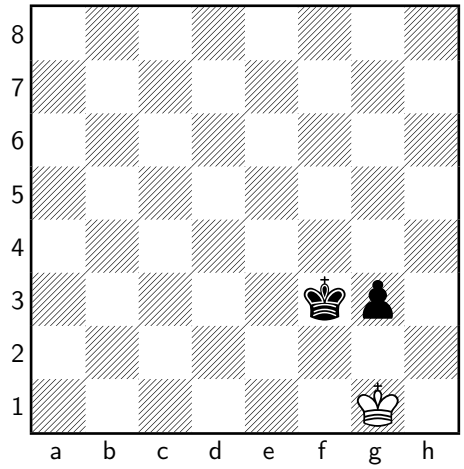

Figure 2.1: The *state of the world* is c_0

It may help some readers to observe the foregoing construction in action. Consider the following "toy model," using the game of chess as an exemplar. To model "actions" in chess using the foregoing framework, let $\mathfrak{C}_{ch}$ be the set of all possible (legal) chess positions. The "set of possible actions" in this context can be identified as the set of *legal chess moves*.[36]

Next, examine what sorts of elements the "set of actions" would contain. Take a particular position for example, call it $c_0 \in \mathfrak{C}_{ch}$, where the only pieces on the board are the White King at g1, the Black King at f3, and a Black Pawn at g3 (*Figure* 2.1). If it is White to move, then there are two legal actions: Kf1 and Kh1. Let us denote the position where the White King is on f1 as c_1 (*Figure* 2.2), and the position where the White King is on h1 as c_2 (*Figure* 2.3).

Now, let A_W represent the set of possible White actions. It follows then that $\langle c_0, \{\langle 1, c_1\rangle\}\rangle \in A_W$ and $\langle c_0, \{\langle 1, c_2\rangle\}\rangle \in A_W$. And there exists no x such that $\langle c_0, \{x\}\rangle \in A_W$ and $x \neq \langle 1, c_1\rangle$ and $x \neq \langle 1, c_2\rangle$. Continuing in a similar fashion for all possible legal chess positions, we may construct a set to represent all possible White moves in all possible positions (i.e., the set A_W). Of course, the legal moves for Black will differ (Black cannot, after all, move the White pieces). And so there will be a distinct set of possible actions $A_B \in \mathfrak{A}(\mathfrak{C}_{ch})$, which identifies every legal *Black* move in every possible position. And the set of all possible actions in chess

[36] Of course moves are reductive lotteries where the outcome is always certain (i.e., if the player moves a pawn forward one square, then it will transform the position to one where the pawn is advanced one square with total certainty). There is zero probability that, upon moving the pawn, the position can be anything else.

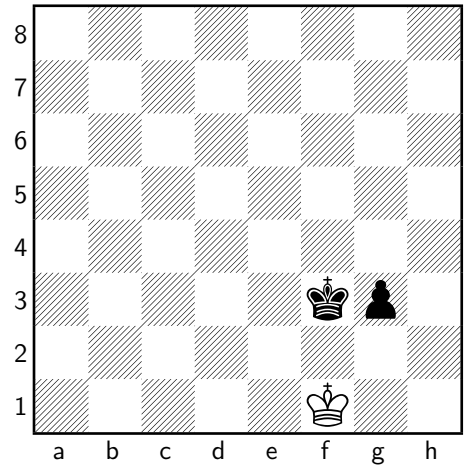

Figure 2.2: The *state of the world* is c_1

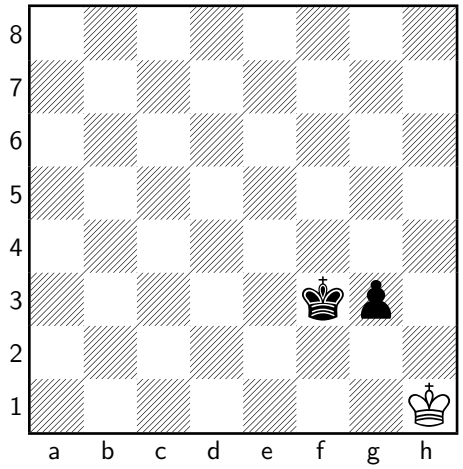

Figure 2.3: The *state of the world* is c_2

may be identified as $A_{ch} = A_W \cup A_B$.

Certainly there exist many other formulations $A_X, A_Y, \ldots \in \mathfrak{A}(\mathfrak{C}_{ch})$, which contain different *sets of possible actions* (for example, chess variants like "suicide chess" and "knight relay chess"). These would be distinct variants on the game of chess, which use the same board and pieces as classical chess (i.e., the same $\mathfrak{C}_{ch}$), but which utilize different rules. The set of *all* possible sets of valid actions (i.e., the set of all variants of chess), $A_W, A_B, A_X, A_Y, \ldots$, given an 8×8 board and the standard set of pieces, is denoted $\mathfrak{A}(\mathfrak{C}_{ch})$.

2.2.3 Hierarchy of Decision-Making Procedures

Next, it will be necessary to construct a hierarchy of decision-making procedures. Let us begin by defining formally what a "decision-making procedure" is. A decision-making procedure is simply a mechanism for deciding among alternative choices.

For example, the set $\{\langle c_0, S_0\rangle, \langle c_1, S_1\rangle, \langle c_2, S_2\rangle\}$ represents one possible decision-making procedure, where $c_0 \neq c_1 \neq c_2$ if $S_0 \neq S_1 \neq S_2$. This set may be interpreted as extensionally equivalent to the conditional imperative: "If the state of the world is c_i, then choose the action represented by $\langle c_i, S_i\rangle$ for $i \in \{0, 1, 2\}$." Note that this representation abstracts away the psychological mechanisms which go into the production of the choice. The psychological aspect of decision-making will be captured by the assignment of decision costs below (Definition 2.2.12). For present purposes, let any decision-making procedure be identified by its inputs and outputs.

In the simplest case, decision-making procedures output sets of actions (e.g., if $S_0, S_1, S_2 \in A \in \mathfrak{A}(\mathfrak{C})$ in the example immediately above). However, decision-making procedures can also have a *nested* structure. In treating the decision *how to decide*, the decision-maker employs a *second order* decision-making procedure. The output of the second order decision-making procedure is not an action, but rather a first order decision-making procedure.

To account for these higher order decision-making procedures, let us construct a hierarchy of decision-making procedures inductively, beginning with the base case of decision-making procedures which output sets of actions only (call the set L^0). Let us then define L^1 decision-making procedures, which select *either* actions *or* L^0 decision-making procedures as its outputs; and L^2 decision-making procedures, which select actions, L^0 decision-making procedures, *or* L^1 decision-making procedures as its outputs; and so forth. The set of decision-making procedures L^i may be thusly defined for any arbitrarily large i, such that the output of any decision-making procedure in L^i may be any member of L^j, such that $0 \leq j < i$.

For convenience, I first define two "conversion functions." These are simply technical tools which will be used in the construction of the hierarchy.

Definition 2.2.4 (Conversion Function I)**.**

a. Given any set of possible conditions $\mathfrak{C}$, for any $A \in \mathfrak{A}(\mathfrak{C})$, and any action $a = \langle c, S\rangle \in A$,[37] *let $\sigma(\langle c, S\rangle) = \langle c, \langle c, S\rangle\rangle$.*

b. And for any set of actions $y \subseteq A \in \mathfrak{A}$, let the "first conversion function" be defined by $h_0(y) = \bigcup_{x \in y}\{\sigma(x)\}$.

The first conversion function inputs a set of possible actions y and outputs a set, such that for all $x \in y$ the output of $h_0(y)$ is the union of all $\{\sigma(x)\}$. For example:

$$h_0(\{\langle c_0, S_0\rangle, \langle c_1, S_1\rangle, \langle c_2, S_2\rangle\}) = \{\langle c_0, \langle c_0, S_0\rangle\rangle, \langle c_1, \langle c_1, S_1\rangle\rangle, \langle c_2, \langle c_2, S_2\rangle\rangle\}$$

Definition 2.2.5 (Conversion Function II)**.**

a. For any function f and any object x, let $\tau(\langle f, x\rangle) = \{\langle x, \langle f, x\rangle\rangle \mid x \in Dom(f)\}$.

b. And for any set of ordered pairs y, such that the first member of every member of y is a function, let the "second conversion function" be defined by $h_1(y) = \bigcup_{z \in y} \tau(z)$.

[37] Definition 2.2.3 guarantees that there will exist some set S, for which $\langle c, S\rangle \in A$, for all $c \in \mathfrak{C}$ and $A \in \mathfrak{A}(\mathfrak{C})$.

The second conversion function inputs a set of ordered pairs, the first member of each being a function. It outputs a set of ordered pairs, such that every element in the domain of every function maps to that function. For example, if $X = \{\langle f, i\rangle, \langle f', j\rangle\}$, and $f = \{\langle c_0, S_0\rangle, \langle c_1, S_1\rangle\}$ and $f' = \{\langle c_2, S_2\rangle, \langle c_3, S_3\rangle\}$, then:

$$h_1(X) = \{\langle c_0, \langle f, i\rangle\rangle, \langle c_1, \langle f, i\rangle\rangle, \langle c_2, \langle f', j\rangle\rangle, \langle c_3, \langle f', j\rangle\rangle\}$$

Definition 2.2.6 (Hierarchy of Decision-Making Procedures)**.** *Given a set of possible actions $A \in \mathfrak{A}(\mathfrak{C})$ and indices $j \in Ord$, and index set I, the set of decision-making procedures is defined hierarchically by the formula:*

$$L^j(A) = \bigcup_{g_k \in G(x)} ((\mathcal{P}(Img(g_k)) \setminus \{\varnothing\}) \times I)$$

$$\textit{Where } x \textit{ is given by: } x = \begin{cases} h_0(A)/\sim & \textit{if } j = 0 \\ \left(\bigcup_{k<j} h_1(L^k(A)) \cup h_0(A)\right)/\sim & \textit{if } j \geq 1 \end{cases}$$

The foregoing definition describes the sets of decision-making procedures, given some $\mathfrak{C}$ and $A \in \mathfrak{A}(\mathfrak{C})$. The sets are defined cumulatively. To simplify notation, I will omit the argument when it is clear from context. I.e., $\ulcorner L^j \urcorner$ refers to $L^j(A)$ when A is understood from context. In the base case, L^0 is defined as the set of all decision-making procedures which output actions and only actions.

Next, L^1 is the set of decision-making procedures which output either actions *or* members of L^0. These are represented as sets, outputting either actions, or members of L^0, again with the restriction that every antecedent condition (triggering *either* actions or subordinate decision-making procedures from L^0) must be unique in each L^1 decision-making procedure.

For example, where $a_0, a_1 \in A \in \mathfrak{A}$ and $y_0, y_1 \in L^0$, one possible decision-making procedure in L^1 is $x = \langle\{\langle c_0, a_0\rangle, \langle c_1, a_1\rangle, \langle c_2, y_0\rangle, \langle c_3, y_1\rangle\}, 0\rangle \in L^1$. The interpretation is straightforward. If a decision-maker were to employ decision-making procedure x, given some state of the world c_k, then in case $k \in \{0, 1\}$, he will choose action a_k. And in case $k \in \{2, 3\}$, then if $y_k = \langle S, j\rangle$ there will exist some $\langle c_k, a_n\rangle \in S$, and he will choose action a_n. Otherwise, x is undefined for conditions where $k \notin \{0, 1, 2, 3\}$.

The second element of every decision-making procedure is simply an identifying marker, which allows multiple extensionally equivalent decision-making procedures to be defined within a model.

Continuing on inductively, L^j may be defined for any arbitrarily large $j \in Ord$. Any ordered pair $\langle f, i\rangle$ is a "decision-making procedure" iff there exists some L^n such that $\langle f, i\rangle \in L^n$. Let $\mathfrak{L}(A)$ refer to the set of all decision-making procedures.

Definition 2.2.7 (Set of All Decision-Making Procedures)**.** *Given some* $A \in \mathfrak{A}(\mathfrak{C})$*, let the set of all decision-making procedures* $\mathfrak{L}(A) = \bigcup_{i \in Ord} L^i(A)$*.*

Lemma 2.2.0.1 (Inclusion)**.** *For any* $A \in \mathfrak{A}(\mathfrak{C})$ *indices* $i, j \in Ord$*, and index set* K*, such that* $i \leq j$*,* $L^i(A) \subseteq L^j(A)$*.*

Proof. First observe the general proposition that for any sets S_0 and S_1, if $S_0 \subseteq S_1$ then for every choice function $g \in G(S_0/\sim)$ there will exist some corresponding $g' \in G(S_1/\sim)$ such that $Img(g) \subseteq Img(g')$ (this follows trivially from Definition 2.2.2). And it follows that:

$$\bigcup_{g_i \in G(S_0/\sim)} (\mathcal{P}(Img(g_i) \setminus \{\varnothing\}) \times K) \subseteq \bigcup_{g_j \in G(S_1/\sim)} (\mathcal{P}(Img(g_j) \setminus \{\varnothing\}) \times K) \tag{2.1}$$

Let S_0 and S_1 be defined by:

$$S_0 = \begin{cases} h_0(A) & \text{if } i = 0 \\ \left(\bigcup_{k<i} h_1(L^k(A)) \cup h_0(A)\right) & \text{if } i \geq 1 \end{cases}$$

$$S_1 = \begin{cases} h_0(A) & \text{if } j = 0 \\ \left(\bigcup_{k<j} h_1(L^k(A)) \cup h_0(A)\right) & \text{if } j \geq 1 \end{cases}$$

Thus, if $S_0 \subseteq S_1$, then it follows that $L^i \subseteq L^j$ for all $i \leq j$ (Definition 2.2.6). And it follows immediately from $i \leq j$ that $S_0 \in S_1$. Therefore, $L^i \subseteq L^j$.

□

Definition 2.2.8 (Transformation Function)**.** *Let the function* $T : \{L^i \mid i \in Ord\} \to \{L^i \mid i \in Ord\}$ *be defined by:* $T(L^i) = L^{i+1}$*.*

The transformation function is helpful to proving the existence of fixed points in the results below.

Abbreviation 2.2.8.1. *For any index* $i \in I$*, any* $A \in \mathfrak{A}(\mathfrak{C})$*, and any decision-making procedure* $\langle f, i\rangle \in \mathfrak{L}(A)$*, let* $\ulcorner f_i \urcorner$ *be an alternative name for* $\langle f, i\rangle$*.*

Theorem 2.2.1 (Fixed Point)**.** *Given some* $A \in \mathfrak{A}(\mathfrak{C})$*, the transformation function* T *has a fixed point in* $\mathfrak{L}(A)$*. That is,* $\exists n\,(T(L^n) = L^n)$*.*

Proof. Lemma 2.2.0.1 establishes that for any $i, j \in Ord$, if $i \leq j$ then $L^i \subseteq L^j$. Observe further that it is the case that $T(L^i) \subseteq T(L^j)$ (again, assuming $i \leq j$) (Definitions 2.2.6, 2.2.8, Lemma 2.2.0.1). Ergo, T is monotonic. And therefore, the existence of a fixed point follows directly from the Knaster-Tarski theorem. □

Definition 2.2.9 (Rank)**.** *Given any set of possible conditions* $\mathfrak{C}$ *and some* $A \in \mathfrak{A}(\mathfrak{C})$*, let* $\lambda = \{\langle f_i, n\rangle \mid \forall x\,(f_i \in L^n \wedge (x < n \implies f_i \notin L^x))\}$*.*

The "rank" of a decision-making procedure is simply the lowest level of the hierarchy defined by Definition 2.2.6 of which that decision-making procedure is a member. In other words, if $\lambda(f_n) = m$, then $f_n \in L^m$ and $f_n \notin L^k$ for all $k < m$. In words, let us thus refer to f_n as being "of rank $m+1$" or "$(m+1)^{\text{th}}$ order" iff $\lambda(f_n) = m$. For example, a decision-making procedure which outputs actions and only actions is a member of L^0, which means it is "first order," or "of rank 1."

2.2.4 Meta-Utility

Definition 2.2.10 (Simple Utility)**.** *For any set of possible conditions* $\mathfrak{C}$*, let* $U(\mathfrak{C}) = \mathfrak{C} \to [0,1]$ *denote the function space of "simple utility functions"* $u_0, u_1, \ldots \in U(\mathfrak{C})$*.*

A simple utility function simply assigns some real number in the interval $[0,1]$ to each state of the world.

Definition 2.2.11 (Simple Expected Utility Functions)**.** *Given any set of possible conditions* $\mathfrak{C}$ *and any* $A \in \mathfrak{A}(\mathfrak{C})$*, let the function* $EU : (A \times U) \to \mathbb{R}$ *be defined by:*

$$EU(\langle c, S\rangle, u) = \sum_{\langle r,c\rangle \in S} ru(c)$$

Definition 2.2.12 (Decision Cost Assignments)**.** *Given any set of possible conditions* $\mathfrak{C}$*, some index set* I*, and any* $A \in \mathfrak{A}(\mathfrak{C})$*, let* $D(\mathfrak{C}, A) = \bigcup_{i \in I}((\mathfrak{C} \times \{f_j \mid f_j \in \mathfrak{L} \wedge \lambda(f_j) = i\}) \to [0, 2^{i+1}])$ *denote the function space of "decision cost assignments"* $d_0, d_1, \ldots \in D(\mathfrak{C}, A)$*.*

A decision cost assignment is a function which assigns each decision-making procedure a numerical value signifying its cost. Importantly, the value assigned to a decision-making procedure is limited by the procedure's rank. For example, a first order decision-making procedure can have a decision cost in the interval $[0,2]$. A second order decision-making procedure can have a decision cost in the interval $[0,4]$. And for any f_n such that $\lambda(f_n) = m$, the decision cost can be in the range $[0, 2^m]$.

The reason why, for any decision-making procedure f_i, the maximum assignable value of $d(c, f_i)$ increases as $\lambda(f_i)$ increases is to allow decision costs to reorder potential behavior completely at each level in the hierarchy. Consider the maximum conceivable impact decision costs can have on decision-making. The most dramatic effect would be to reverse a decision-maker's behavior, so that he behaved *as if* his

most preferred state of the world were his least preferred state of the world. Thus, if in some model, $u(c_n) = 0$ and $u(c_m) = 1$, then in the extremum case, decision assignment would have to be capable of reordering his behavior so that he behaved as if $u(c_n) > u(c_m)$.

Definition 2.2.13 (Model)**.** *Given any set of possible conditions* $\mathfrak{C}$*, let* $\mathfrak{M}(\mathfrak{C}) = \{\langle A, u, d\rangle \mid A \in \mathfrak{A}(\mathfrak{C}) \wedge u \in U(\mathfrak{C}) \wedge d \in D(\mathfrak{C}, A)\}$ *denote the set of all models* $M_0, M_1, \ldots \in \mathfrak{M}(\mathfrak{C})$*.*

A model represents all the relevant variables in the structure of decision-making, given a set of possible states of the world: the set of all possible actions, a utility function, and a decision cost assignment.

Definition 2.2.14 (Meta-Utility Functions)**.** *Given any set of possible conditions* $\mathfrak{C}$*, let* $\pi(c) \in [0, 1]$ *represent the probability that some* $c \in \mathfrak{C}$ *occurs.*[38] *Next let* Q *be the set of all ordered triples* $\langle c_i, f_j, M_k\rangle$*, such that* $M_k = \langle A, u, d\rangle \in \mathfrak{M}(\mathfrak{C})$ *and* $f_j = \langle f, j\rangle \in \mathfrak{L}(A)$ *and* $c_i \in Dom(f)$*. More formally:*

$$Q = \{\langle c_i, f_j, M_k\rangle \mid \exists A, u, d\,(M_k = \langle A, u, d\rangle \in \mathfrak{M}(\mathfrak{C}) \wedge f_j \in \mathfrak{L}(A) \wedge c_i \in Dom(f))\}$$

And let "meta-utility functions" $\mu : Q \to \mathbb{R}$ *be defined by:*

$$\mu(c, f_n, M) = \begin{cases} \pi(c)\,(EU(f(c), u) - d(c, f_n)) & \text{if } f(c) \in A \\ \mu(c, f(c), M) - \pi(c)d(c, f_n) & \text{if } f(c) \notin A \end{cases}$$

Intuitively, the meta-utility function $\mu(c, f_n, M)$ returns something *like* expected utility. It is the "expected utility" of using a particular decision-making procedure f_n, given some state of the world $c \in \mathfrak{C}$ in some model $M \in \mathfrak{M}(\mathfrak{C})$. For first order decision-procedures, this reduces to the simple expected utility of choosing the action determined by f_n less the decision cost of f_n, all discounted by the probability that c obtains. For higher order decision-making, the calculation is essentially the same, except that each higher order decision-making procedure imposes an additional expected decision cost.

Definition 2.2.15 (Procedure Utility)**.** *Given any set of possible conditions* $\mathfrak{C}$*, let the "procedure utility" function* $\rho : \{\langle f_i, M_j\rangle \mid \exists A, u, d\,(\langle A, u, d\rangle = M_j \in \mathfrak{M}(\mathfrak{C}) \wedge f_i \in \mathfrak{L}(A))\} \to \mathbb{R}$*, be defined by:*

$$\rho(f_n, M) = \sum_{c_i \in Dom(f)} \mu(c_i, f_n, M)$$

[38] Note that $\sum_{x \in \mathfrak{C}} \pi(x) = 1$.

The "procedure utility" of a decision-making procedure f_n is the sum of the meta-utility values $f(c)$ for every element c in the domain of f. Intuitively, it is the expected meta-utility of the decision-making procedure for every condition in which it could be deployed.

Definition 2.2.16 (Global Decision-Making Procedures)**.** *Given any set of possible conditions $\mathfrak{C}$, any set of possible actions $A \in \mathfrak{A}(\mathfrak{C})$, and any set of procedures S, let $\Gamma : \mathcal{P}(\mathfrak{L}(A)) \to \mathcal{P}(\mathfrak{L}(A))$ return the subset of S, such that the domain of every decision-making procedure is equal to $\mathfrak{C}$. More formally: $\Gamma(S) = \{f_i \in S \mid Dom(f) = \mathfrak{C}\}$.*

Intuitively, a "global decision-making procedure" is a decision-making procedure which is defined for every possible state of the world. Global decision-making procedures are significant, because they can represent not only a method of determining a choice, but a decision-maker's behavior *in toto*.

Definition 2.2.17 (S-Rational Procedures)**.** *Given any set of possible conditions $\mathfrak{C}$, and any model $M = \langle A, u, d\rangle \in \mathfrak{M}(\mathfrak{C})$, and any set of decision-making procedures S, let $R : (\mathcal{P}(\mathfrak{L}(A)) \times \mathfrak{M}) \to \mathcal{P}(\mathfrak{L}(A))$ return the subset of $\Gamma(S)$, defined by:*

$$R(S, M) = \{f_i \mid \forall f_j \, (f_i \in \Gamma(S) \wedge (f_j \in \Gamma(S) \implies (\rho(f_i, M) \geq \rho(f_j, M))))\}$$

Intuitively, $R(L^n, M)$ returns the set of decision-making procedures in L^n such that every member of $R(L^n, M)$ returns a procedure utility value greater than or equal to any other global decision-making procedure in L^n. That is, $R(L^n, M)$ is the set of "L^n-rational decision-making procedures."

Theorem 2.2.2 (Ratcheting)**.** *For any $M = \langle A, u, d\rangle \in \mathfrak{M}(\mathfrak{C})$ and $f_i \in \Gamma(L^k(A))$ and $f'_j \in R(L^l(A), M)$ such that $k \leq l$, it must be the case that $\rho(f_i, M) \leq \rho(f'_j, M)$.*

Proof. Suppose for *reductio* that $\rho(f_i, M) > \rho(f'_j, M)$. Because $k \leq l$, it follows from Lemma 2.2.0.1 that $L^k \subseteq L^l$, and therefore that $f_i \in L^l$. Now if $\rho(f_i, M) > \rho(f'_j, M)$, then this contradicts the assumption that $f'_j \in R(L^l, M)$ (Definitions 2.2.15, 2.2.17).

Thus, it cannot be the case that $\rho(f_i, M) > \rho(f'_j, M)$, and therefore it must be the case that $\rho(f_i, M) \leq \rho(f'_j, M)$. □

The ratcheting theorem implies that the procedure utility of any L^i-rational procedure must be greater than of equal to the procedure utility of any L^j-rational procedure if $i > j$. In other words, the procedure utility of L^i-rational procedures cannot decrease as i increases.

Corollary 2.2.2.1 (Boundedness)**.** *For any $M = \langle A, u, d\rangle \in \mathfrak{M}(\mathfrak{C})$, any index n, and any $f_i \in R(L^n, M)$, the procedure utility of f_i must be bounded, such that $-2 \leq \rho(f_i, M) \leq 1$.*

Proof. Let us first prove the upper bound $\rho(f_i, M) \leq 1$, for all $f_i \in L^n(A)$ and $M \in \mathfrak{M}(\mathfrak{C})$. Consider any $a = \langle c, S\rangle \in A \in \mathfrak{A}(\mathfrak{C})$. Since $\sum_{\langle r,c\rangle \in S} r \leq 1$ (Definition 2.2.3), and $u(c) \leq 1$ for all $c \in \mathfrak{C}$ (Definition 2.2.10), it follows that $EU(a, u) \leq 1$ for all $u \in U$ (Definition 2.2.11). Because the only positive term in the construction of $\mu(c, f_i, M)$ is $EU(a, u)$ (Definition 2.2.14), and $\sum_{c_j \in \mathfrak{C}} \pi(c_j) \leq 1$, it follows that $\rho(f_i, M) \leq \sum_{c_j \in Dom(f)} \pi(c_j)\, EU(\theta(c_j, f(c_j)), u) \leq 1$ (Definition 2.2.15).

To prove the lower bound $-2 \leq \rho(f_i, M)$, let us start with first order procedures and generalize inductively. Observe that for any $f_i \in R(L^0(A), M)$, for every $a \in Img(f_i)$, $EU(a, u) \geq 0$ (Definition 2.2.10, 2.2.11), and its decision assignment is bounded $d(c, f_i) \leq 2$ because $\lambda(f_i) = 0$ (Definition 2.2.12). Thus, in the extremum case that $d(c, f_i) = 2$ and $EU(f(c), u) = 0$, the meta-utility will be bounded $EU(f(c), u) - d(c, f_i) \geq -2$. Therefore, since $\sum_{c_j \in Dom(f)} \pi(c_j) \leq 1$, it follows that $-2 \leq \sum_{c_j \in Dom(f)} \pi(c_j)(EU(f(c_j), u) - d(c_j, f_i)) = \rho(f_i, M)$ (Definition 2.2.15). Now, having established that for any $f_i \in R(L^0(A), M)$ the procedure utility $\rho(f_i, M) \geq -2$, it follows from Theorem 2.2.2 that for any arbitrarily large k, if $f_i \in R(L^k, M)$ then $-2 \leq \rho(f_i, M)$. □

Definition 2.2.18 (Behavior Function)**.** *Given any set of possible conditions $\mathfrak{C}$, let the set of "behavior functions" $B : \mathfrak{A}(\mathfrak{C}) \to h_0(\bigcup \mathfrak{A}(\mathfrak{C}))$ be defined by:*

$$B(A) = \bigcup_{g_i \in G(A/\sim)} \{Img(g_i)\}$$

Denote particular behavior functions using lowercase $b_0, b_1, \ldots \in B$. A behavior function simply maps states of the world to actions, representing the potential behavior of a decision-maker.

Definition 2.2.19 (Reduction Function)**.** *For any $c \in \mathfrak{C}$ and any set of possible actions $A \in \mathfrak{A}(\mathfrak{C})$, let the "reduction function," $\theta : \mathfrak{C} \times \mathfrak{L}(A) \to \mathcal{P}(A)$, be defined by:*

$$\theta(c, f_n) = \begin{cases} f(c) & \text{if } f(c) \in A \\ \theta(c, f(c)) & \text{if } f(c) \notin A \end{cases}$$

The reduction function returns the action determined by a decision-making procedure given some state of the world. For first order decision-making procedures, the

output is trivial. However, for higher order decision-making procedures, the effect of the reduction function is to unnest the hierarchy of decision-making procedures to output an action. For example, if the state of the world is c_0, and decision-making procedure f_0 determines that when c_0 obtains, a decision-maker should use decision-making procedure f_0'; and decision-making procedure f_0' determines that when c_0 obtains, a decision-maker should use decision-making procedure f_0''; and decision-making procedure f_0'' determines that when c_0 obtains, a decision-maker should do a_0, then the "reduction" of f_0 is that ultimately the decision-maker should do a_0.

More formally: suppose that $f_0 = \langle f, 0\rangle \in L^2(A)$. And suppose $f(c_0) = f_0'$. And suppose that $f'(c_0) = f_0''$. And suppose that $f''(c_0) = a_0 \in A$. It follows from the definition of the reduction function therefore that, $\theta(c_0, f_0) = a_0$.

Definition 2.2.20 (Decomposition Function)**.** *Given any set of possible conditions $\mathfrak{C}$, and any set of possible actions $A \in \mathfrak{A}(\mathfrak{C})$, let the "decomposition function," $\delta : \mathfrak{L}(A) \to \mathcal{P}(h_0(A))$, be defined by:*

$$\delta(f_n) = \bigcup_{c_i \in Dom(f)} \{\theta(c_i, f_n)\}$$

The decomposition function gives the set of all conditional behaviors determined by a given decision-making procedure. In other words, it "collapses" any i^{th} order decision-making procedure into a behavior function. For example, the decision-making procedure $x = \langle\{\langle c_0, \langle\{\langle c_0, \langle\{\langle c_0, a_0\rangle, \langle c_1, a_1\rangle\}, 0\rangle\rangle, \langle c_2, a_2\rangle\}, 0\rangle\rangle, \langle c_3, a_3\rangle\}, 0\rangle$ may be decomposed $\delta(x) = \{a_0, a_3\}$.

Definition 2.2.21 (Meta-Rational Procedure)**.** *Given any set of possible conditions $\mathfrak{C}$ and some model $M \in \mathfrak{M}(\mathfrak{C})$, where $L^F(A)$ is the least fixed point of T, let a "meta-rational procedure" in M be any f_i such that $f_i \in R(L^F, M)$.*

Let us signify that some f_m is a meta-rational choice procedure in model M with the superscript notation $\ulcorner f_m^{M*} \urcorner$. And let us say that f_m^{M*} "models" $b \in B(A)$ iff $\delta(f_m^{M*}) = b$.

Theorem 2.2.3 (Finiteness of Meta-Rational Meta-Utility Values)**.** *Given any set of possible conditions $\mathfrak{C}$, and any $c \in \mathfrak{C}$ and any $M \in \mathfrak{M}(\mathfrak{C})$, and any meta-rational procedure f_i^{M*}, the meta-utility of every decision is bounded $-2 \leq \mu(c, f_i^{M*}, M) \leq 1$.*

Proof. This follows directly from Definition 2.2.21 and Corollary 2.2.2.1. □

Theorem 2.2.4 (Representability)**.** *Given any set of possible conditions $\mathfrak{C}$, any $A \in \mathfrak{A}(\mathfrak{C})$, any behavior function $b \in B(A)$, and any index $j \in \mathbb{N}$, there exists an f_n, such that $\delta(f_n) = b$ and $\lambda(f_n) = j$ and $|Img(f)| > 1$.*

Proof. First, observe that for any choice function g, it must be the case that $g \in G(h_0(A)/\sim)$ iff $Img(g) \in B(A)$ (Definition 2.2.18). Next, note that for any x, it must be the case that $x \in \mathcal{P}(x)$; therefore, *a fortiori*, for any choice function g, it must be the case that $Img(g) \in \mathcal{P}(Img(g))$. Consequently, for any function g, and any index $i \in I$, it must be the case that $g \in G(h_0(A)/\sim)$ iff $\langle Img(g), i\rangle \in L^0(A)$ (Definition 2.2.6). And by transitivity, $b \in B(A)$ iff $\langle b, i\rangle \in L^0(A)$. And it follows from Definition 2.2.20 therefore that for any $b \in B(A)$, there exists some $f_0 \in L^0(A)$, such that $\delta(f_0) = b$.

Next, if $|\mathfrak{C}| > 1$ and there exists some f_0', such that $\lambda(f_0') = k$ and $\delta(f_0') = b$, then there must exist functions f_0'' and f_0''', such that $\lambda(f_0'') = k$ or $\lambda(f_0''') = k$, and $f'' \cup f''' = f'$. Note that $\delta(f_0'') \cup \delta(f_0''') = \delta(f_0')$ (Definition 2.2.20). Now, define $f_0^\dagger = \langle\{(Dom(f'') \times \{f_0''\}) \cup (Dom(f''') \times \{f_0'''\})\}, 0\rangle$. Since $\lambda(f_0'') = k$ or $\lambda(f_0''') = k$, it must be the case that $\lambda(f_0^\dagger) = k + 1$ (Definition 2.2.6). And $b = \delta(f_0') = \delta(f_0'') \cup \delta(f_0''') = \delta(f_0^\dagger)$ (Definition 2.2.20). And by construction, $\left|Img(f^\dagger)\right| = 2$ (trivially, this value can be extended to any $x > |\mathfrak{C}|$ by analogous construction).

Now, since we have already established that there exists an f_0, such that $\lambda(f_0) = 0$ and $\delta(f_0) = b$ for each $b \in B$, and since we have also established that if there exists an f_0', such that $\lambda(f_0') = k$ and $\delta(f_0') = b$, then there must exist an $f_0^\dagger$, such that $\lambda(f_0^\dagger) = k + 1$ and $\delta(f_0^\dagger) = b$, it follows by induction that for any j there exists an $f_0^{\dagger\dagger}$, such that $\delta(f_0^{\dagger\dagger}) = b$ and $\lambda(f_0^{\dagger\dagger}) = j$. Note that the identifier index is irrelevant for present purposes.[39] □

Theorem 2.2.5 (Existence of a Meta-Rational Representation)**.** *Given any set of possible conditions $\mathfrak{C}$, any $A \in \mathfrak{A}(\mathfrak{C})$, and any behavior function $b \in B(A)$, there exist infinitely many models $M = \langle A, u, d\rangle \in \mathfrak{M}(\mathfrak{C})$, for which there exists some f_n^{M*}, such that $\delta(f_n^{M*}) = b$.*

Proof. Theorem 2.2.1 establishes the existence of a fixed point, L^F. Theorem 2.2.4 establishes that for any $b \in B$, there must exist some $f_m \in \mathfrak{L}$, such that $\delta(f_m) = b$. What remains is to show that f_m is a meta-rational procedure for some $M \in \mathfrak{M}$.

This can be demonstrated trivially with an *ad hoc* construction. Let u be any member of $U(\mathfrak{C})$. We will want to define a decision cost assignment $\overline{d_m}$, such that $f_m \in R(L^F, M = \langle A, u, \overline{d_m}\rangle)$. First take any arbitrary decision cost assignment $d_m \in D(\mathfrak{C}, A)$. Now, define $\overline{d_m}$ by:

$$\overline{d_m}(c, f_i') = \begin{cases} d_m(c, f_i') & \text{if } \lambda(f_i') < \lambda(f_n) \\ 2^{\lambda(f_i')+1} & \text{if } \lambda(f_i') \geq \lambda(f_n) \text{ and } f_i' \neq f_n \\ 0 & \text{if } \lambda(f_i') \geq \lambda(f_n) \text{ and } f_i' = f_n \end{cases}$$

[39] For any f_i and f_j, $\delta(f_i) = \delta(f_j)$.

Observe that $\overline{d_m} \in D(\mathfrak{C}, A)$ if $d_m \in D(\mathfrak{C}, A)$ (Definition 2.2.12). Also, it must be the case that $f_n \in R(L^{\lambda(f_n)}, M)$ by construction. And now, because every decision-making procedure $f_j'' \neq f_n$ such that $\lambda(f_j'') \geq \lambda(f_n)$ is assigned maximum costs, i.e., $2^{\lambda(f_j'')+1}$, it must have a procedure utility less than f_n; and the greatest meta-utility value of any decision-making procedure f_k''' such that $\lambda(f_k''') < \lambda(f_n)$ would be $\sum_{l=1}^{\lambda(f_n)} 2^l < 2^{\lambda(f_n)+1}$. Therefore it must be the case that $f_n \in R(L^F, M)$. □

2.3 Interpretation and Implications

2.3.1 Informal Restatement

It is worth restating the foregoing formal model and its results in words. The concept of "global decision-making procedures" is the most intuitive starting point from which to access the theory. A decision-making procedure is "global" if it accounts for all the decisions that an individual might encounter, specifying the choices he *would* make in each circumstance.[40]

Imagine an exceedingly meticulous decision-maker planning for contingencies. Determined that he should never face a decision for which he is unprepared, let us suppose he maps out a plan of action for every conceivable state of the world. He will *do* x_0 if he encounters state of the world c_0; he will *do* x_1 if he encounters state of the world c_1; and so forth, for every possible state of the world. The conjunction of all the potential plans he would select form a "grand plan of action." That grand plan of action is a global decision-making procedure. It is "global" because it accounts for every possible contingency. Note that a global decision-making procedure perfectly represents an individual's potential future behavior.[41] Also note that I am *not* claiming that any real individual could (much less does) formulate such a grand plan of action—the hypothetical is meant to be illustrative.

Now let us consider what kinds of things an x_i in the schema "if c_i then do x_i" could be. In the meticulous decision-maker's grand plan of action, x_i could simply be *to do* some action; but in other cases, x_i may be a *subordinate plan* for selecting an action, and in other cases still, a subordinate plan for selecting a subordinate plan for selecting an action.

For example, the meticulous decision-maker could adopt the general rule: that if he is feeling lethargic and sees a nearby coffee shop, then he will purchase a caffè

[40] This is characterized formally in Definition 2.2.16, *supra*.

[41] By "perfect," I mean that it predicts his behavior without error, so long as he does not deviate from the plan. Of course, if he *does* deviate from the plan, then we would not call the plan a valid representation.

macchiato. Of course, his plan could be more nuanced. He could make it a rule that he will order a cappuccino if it is earlier than 10:00 a.m., and a macchiato otherwise; or he could order a caffè americano if he is in a contemplative mood, and a macchiato if he is in a whimsical spirit. And of course the triggering conditions need not be stated in general terms. In other words, it need not be a *rule*. His plan could be a particularized enumeration of actions: ordering a macchiato at 3:23 p.m. on July 10, 2018; a caffè corretto at 7:32 p.m. on July 11, 2018; and so forth, for every future coffee purchase he will make in his life.

However, a decision-making procedure need not *directly* prescribe any particular action at all. Decision-making procedures can boot the decision to another subordinate decision-making procedure. For example, the meticulous decision-maker's plan of action might prescribe: that if he is feeling lethargic, then he should locate the nearest coffee shop, determine what the most popular beverage happens to be at that establishment, and order *whatever that is*. The action is not fixed in this case. Rather, it is the outcome of another subordinate plan. And of course, the subordinate plan could boot the problem to a more subordinate plan.

Now, there are two objectives for the meta-model. It must first establish that for any set of potential actions which a decision-maker would perform, it is possible to construct some global decision-making procedure which prescribes all those potential actions and only those potential actions (or boots the decision to subordinate decision-making procedures which prescribe all those potential actions and only those potential actions). Second, the theory must establish that there exists some model, which characterizes that representative global decision-making procedure as some sort of utility-maximization.

In order to see how the meta-model accomplishes this, we need to specify the meaning of "meta-utility" and "model." Let us start with meta-utility. The meta-utility of a decision is simply the expected utility of a selected action, less the aggregate expected decision costs associated with selecting that choice. It is important to recognize that meta-utility values are downward looking in the hierarchy. An n^{th} order decision-making procedure must cash out in some $(n-1)^{\text{th}}$ order decision-making procedure, which must cash out in some $(n-2)^{\text{th}}$ order decision-making procedure, and so forth until an action is selected by some first order decision-making procedure. The meta-utility of the n^{th} order procedure includes the expected utility of the action and the expected decision costs of every $(m \leq n)^{\text{th}}$ order procedure. But the meta-utility of the $(n+1)^{\text{th}}$ order decision-making procedure which selected the n^{th} order decision-making procedure is not included in the meta-utility of the n^{th} order procedure.

Let us consider this in terms of our meticulous decision-maker's plan of action.

Suppose he is lapsing toward a torpid state, and he believes that a coffee would revive him. He can select from among infinitely many plans—he might seek out the nearest coffee shop and order whatever is most popular there; he might seek out a particular familiar coffee shop and order his favorite beverage; he might weigh the purchase of coffee against taking a nap; or he might undertake a rational deliberation to determine the optimal action.

Suppose he chooses to drink a macchiato at the nearest coffee shop. The meta-utility of his "lethargy plan" is the expected utility of consuming the macchiato, less the expected decision cost of the lethargy plan, less the expected decision costs of all the subordinate plans of action which led to the consumption of the macchiato. Since the meta-utility is always downward looking in the hierarchy, this does not include the meta-utility of the "grand plan" which selected the "lethargy plan."

The principal contention of the meta-rational theory is that individuals are meta-utility maximizers. Of course, this raises the question: meta-utility-maximization *at what order* of decision-making? Since decision-makers must decide how to decide how to decide, etc., meta-utility values would seem to decrease as order increases, as ever more decision costs are heaped on at each additional order of decision-making, with no stopping point. This is the infinite regress problem described above.

Theorem 2.2.1 establishes that there exists a least fixed point in the infinite hierarchy of decision-making—a point at which the hierarchy reaches a limit. There being no principled finite stopping point to the chain of questions—how the decision-maker decides to decide to decide, etc.—the theory identifies "meta-rational procedures" as those which maximize meta-utility irrespective of order. The existence of the fixed point establishes that there exists a determinate (albeit infinitely large) set of possible decision-making procedures, from which the decision-maker may select a maximizing alternative. A "meta-rational procedure"—i.e., the best grand plan of action—is whatever global decision-making procedure returns the greatest meta-utility value when the hierarchy is extended to the fixed point. In other words, it is a global decision-making procedure in the fixed point (which includes all orders of decision-making below the fixed point), which maximizes meta-utility. Note that this need not be a decision-making procedure at the highest order of decision-making. It *could* be some first order decision-making procedure. It is whatever global decision-making procedure at *any* order of decision-making maximizes meta-utility.

Implicitly, the construction operates *as if* the decision-maker were choosing *both* the optimal decision-making procedure *and* the optimal order of decision-making. It endogenizes the question what order of decision-making from which the meticulous decision-maker should choose his grand plan of action.

The conceit is that the meticulous decision-maker, when adopting a grand plan

of action, surveys the vast realm of all possible plans, and adopts whichever global decision-making procedure returns the greatest meta-utility value. All of the subordinate plans he chooses are implicit in the grand plan. The meta-utility maximizing grand plan could be some deeply nested decision-making procedure; or it could be a first order decision-making procedure. The point is that whatever grand plan the meticulous decision-maker adopts, it is the meta-rational plan.

Yet the persistent skeptic will object that if the meticulous decision-maker *chooses* the meta-utility maximizing grand plan, then this implies that he is employing a rational deliberation at some even higher order of decision-making. This is a mistaken interpretation of the "grand plan." It is *technically* mistaken, because the meta-rational procedure was determined at a fixed point. There can be no "grander plan" which selects the "grand plan," or else the "grander plan" *would have been* the "grand plan" selected from all the plans in the fixed point. It is *conceptually* mistaken, because the decision-maker does not *choose* the meta-rational procedure—he *is* the meta-rational procedure. It cannot be turtles all the way down. A hierarchy of plans selecting plans selecting plans, etc., cannot describe which plan is ultimately chosen. At the fundament, there must be a root decision-making procedure which determines which plans are in fact adopted.

Note that the meta-utility-maximization does not imply that the decision-maker has employed a rational deliberation. His behavior can be rational without his having undertaken a rational deliberation. Indeed, the idempotent hierarchy problem guarantees that he is *not* adopting rational deliberation at the fundamental level. Rather, the theory claims that meta-rationality is a *property* of whatever grand plan he adopts. The grand plan need not be—nor would ever realistically be—to deliberate rationally. The grand plan is not *to do a maximization*, but rather has the property *of being maximizing*, whatever it prescribes the decision-maker do.

This is a subtle but important point. A grand plan of action is not *really* so "grand," if it is still subordinate to some more fundamental plan of action. If decision-makers are ever to act, there *must* be a starting point, at which the decision-maker's behavior is not chosen but simply *done*. It is important to recognize that this observation is *not the same* as positing an *ad hoc* starting point. The theory only claims that there must be a starting point—some procedure which is not chosen, but which simply *is* what the decision-maker does. No specific procedure—certainly not rational deliberation—is assumed in the meta-model.

Finally, the numerical values representing the decision-maker's expected utility and decision costs are *models*. Because *any* values can be assigned, infinitely many models can be constructed, representing states of the world, utilities, and decision costs. Some models may lack sufficient expressive capability to describe the decisions

he faces—these would be defective models because they fail to represent decision problems *ab initio*. Other models may fail to determine a meta-rational procedure, because no meta-utility suprema exist in the set of all decision-making procedures (i.e., the least fixed point). Other models may specify meta-utility values, the maximizations of which fail to reflect the decision-maker's behavior in fact—these would be defective models because their meta-rational procedures differ from his actual grand plan of action. Yet among the infinitude of possible models, it *must* be the case that *some* models validly represent the meticulous decision-maker's grand plan as a meta-utility-maximizing global decision-making procedure. The existence of such models is established in Theorem 2.2.5. Moreover, Corollary 2.2.2.1 establishes that the meta-utility of that meta-rational procedure must be some finite value.

In other words, whatever the lethargic decision-maker chooses to do—whether purchasing the house special at a nearby coffee shop, or taking a nap, or hopping on one foot backwards to the post office—there exists a model which can describe that behavior as consistent with a meta-utility-maximization. In other words, there is a meta-rational grand plan of action, which prescribes he do precisely what he does in fact.

Let us return finally to the hypothetical utility-maximizing robot leitmotif. The theory here presented maintains the principal claim of bounded rationality: that a *meta*-utility-maximizing robot *can* replicate all the potential choices of a human being if given two initializing inputs: (i) a utility function, and (ii) a decision cost assignment. Theorem 2.2.5 proves that for any set of potential behaviors, there *must always* exist a set of inputs which would reproduce those potential choices in a meta-utility-maximizing robot. The weak interpretation of the rational actor hypothesis is thus proven true.

2.3.2 Implications

The theoretical payoffs of the meta-rational conception are several. The most important payoff is that the meta-model solves the four problems identified in Section 2.1.3 above.

The infinite regress problem is solved, because the set of all decision-making procedures constructed in Definition 2.2.6 represents a *complete* set containing every conceivable decision-making procedure which could possibly represent a decision-maker's behavior (for a given model). Whatever set of potential behaviors describes the decision-maker's choices, that set will correspond with global decision-making procedures in the fixed point, and there will exist models which characterize a subset of those decision-making procedures as being meta-rational.

The question giving rise to the infinite regress—why the decision-maker would choose *that* decision-making procedure—is answered endogenously. Any decision-making procedure of a higher order than the meta-rational procedure would also have been an element of the set containing all conceivable decision-making procedures. The meta-rational decision-making procedure cannot have been the output of a higher decision-making procedure, or else *that other* decision-making procedure would be the meta-rational procedure. Consequently, it cannot be the case that the meta-rational procedure was "chosen," and therefore the question *why* it was chosen is rendered meaningless. Additionally, Corollary 2.2.2.1 neutralizes the threat of infinite decision costs.

The idempotent hierarchy problem is solved, because decision-makers are not represented as adopting *rational deliberation* at any order of decision-making greater than the first order. Rational deliberation in the meta-model is simply one among many possible decision-making procedures. The meta-rational theory only claims that individuals would undertake a rational deliberation iff rational deliberation were the meta-rational chocie. Of course, rational deliberation may not be meta-rational (indeed, it *cannot* be in fact).

Meta-utility maximization is a *property* of meta-rational procedures. The claim is not that decision-makers *choose* the meta-rational procedure. Rather, the claim is that a model exists which represents the decision-maker's behavior as optimal with respect to meta-utility. This is a nuanced point, which should be distinguished from the "as if" arguments often given for the weak interpretation of classical rationality. The meta-rational answer is more substantive. Once again, the meta-rational decision-maker *cannot* have employed rational deliberation to choose the meta-rational procedure, because the meta-rational procedure cannot have been the output of any other decision-making procedure. The theory simply *does not claim* that meta-rationality is the product of a rational deliberation. Therefore, there can be no paradox.

The explanatory vacuousness of generalized expected utility theories is also solved, because the meta-rational conception explicitly embeds maximization within the framework of bounded rationality. Note that the meta-rational meta-model *is* a species of generalized expected utility, because in any model of the meta-model, the decision-maker's behavior is characterized as *maximizing* something—specifically, meta-utility (again, this should not be confused with the claim that the behavior is *to maximize*).

And finally, the fragility of prior generalized expected utility theories is circumvented, because Theorem 2.2.5 will hold true for *any* set of possible behaviors. Critics may complain that this "feature" of the meta-rational approach renders the meta-

theory unfalsifiable and therefore unscientific. This grossly misunderstands the point. *Of course* the meta-model is unfalsifiable. That is why it is a *meta*-model and not simply a *model*. It is a model *of models*. It is a theory about how behaviors can be represented. The results of the meta-model are entirely about the models definable in it. And models are abstractions: there is nothing claimed at the meta-theoretical level for empirical data to verify or falsify. It is not even clear what "falsifiability" would mean for a *meta*-theory. Indeed, if falsifiability were taken as an acceptability condition of meta-theories, then the scientific method (and indeed, the principle of falsifiability itself) should likewise be regarded defective. It is worth noting additionally, to answer the misguided Popperian critique directly, that any of the *models* definable in the meta-model would in principle be falsifiable.

In addition to resolving the four problems identified above, the meta-rational approach supplies further benefits still. Methodologically, the economic conventionalist should be pleased that the meta-model grounds classical expected utility analysis. Conventionalists have often justified the use of classical expected utility analysis as "assuming away frictions." However, this excuse—without saying anything more—is question-begging. The experimental and empirical observations of behavioralists suggest systematic effects far more serious than mere "frictions." They imply fundamentally different principles in human behavior than expected utility theory assumes. For the justification to be persuasive, the "frictions" excuse requires a more robust account than methodological conventionalists have heretofore provided.

The meta-rational meta-model provides foundation to the claim that apparently non-rational behaviors are indeed the product of frictions. The support is not merely that meta-rationality *resembles* classical rationality in form. Rather, it allows for an explicit characterization of decision costs as "frictions." If we "assume away frictions," as the conventionalists purport to do, then this would specify the set of models with the decision cost assignment $d_n : (\mathfrak{C} \times \mathfrak{L}) \to \{0\}$, assigning zero decision costs to every decision-making procedure in $\mathfrak{L}$. Now observe that the meta-rational procedure in any model containing d_n would simply reduce to expected utility maximization. In other words, expected utility theory *simply is* meta-rational choice theory with a constant assignment of decision costs.[42]

Next, theorists of bounded rationality should be pleased that the meta-model solves the infinite regress and idempotent hierarchy problems. These obstacles had threatened to undermine the essential premises of bounded rationality. But more than this, the meta-model can be used to justify the previously *ad hoc* assumption of

[42]The choice of zero is merely a way of matching the language of the "frictions" narrative. In fact, any constant decision cost assignment would reduce a meta-rational model to an expected utility model.

second order rationality. Observe that all meta-rational models can be "flattened" into a theory of second order decision-making (albeit with some loss of expressivity). More precisely, a function $\beta : \{\langle c, f_i\rangle \mid c \in Dom(f_i) \wedge f_i \in \mathfrak{L}(A)\} \rightarrow (h_0(A) \times \mathbb{R})$ can be constructed, defined by: $\beta(c, f_i) = \langle h_0(\theta(c, f_i)), \mu(c, f_i, M)\rangle$, which collapses every decision-making procedure in an infinite hierarchy of decision-making, along with its corresponding meta-utility, into a set of first order decision-making procedures.

Of course, merely defining β is not sufficient to establish a complete and consistent theory of second-order bounded rationality. However, the remaining work is clearly only a matter of working out the formal details. Consequently, the assumption of costless second order rationality (with no higher orders of decision-making) can be justified by mapping to any traditional two-layer theory of bounded rationality from a meta-rational model. Since the meta-rational model does not make *ad hoc* assumptions about the height of the hierarchy, the two-layer model can be interpreted not be an arbitrary restriction, but rather as a simplification of the meta-rational representation.

Third, theorists of alternative generalized expected utility theories should be pleased that the meta-model can be used to bridge the gap between their theories and bounded rationality. Alternative decision-making models may be subsumed under the meta-rational theory as describing particular decision-making procedures, which meta-rational individuals are "likely to choose" under specified conditions (for example, conditions involving risk or uncertainty).

The meta-rational choice theory developed here is thus an ecumenical conception. The only viewpoint which receives no succor is the menagerie of biases and heuristics, which, as argued in Section 2.1.2, should not have been regarded plausible in the first place.

2.3.3 Strong and Weak Interpretations of Meta-Rationality

Just as conventional conceptions of the rational actor admit both strong and weak interpretations, so too will the meta-rational meta-model admit analogous "strong" and "weak" interpretations.

The "weak" interpretation is simply that there always exists a meta-rational procedure in some model, which represents a decision-maker's behavior. The weak interpretation of the meta-rational thesis is simply that there exist models under which a meta-rational procedure is representative of behavior, and Theorem 2.2.5 establishes that the proposition must be true.

The "strong" interpretation takes utilities and decision costs as *real* phenomena, and by implication implies that meta-utilities are also *real*. The strong interpretation

of the meta-rational thesis is somewhat more appealing than the strong interpretation of expected utility theory for several reasons. First, its psychological claims are more circumscribed. A meta-rational procedure can be *any* psychological process whatever. The strong interpretation of meta-rationality does not specify *how* a decision-maker decides, but merely asserts that however he does decide, that representative decision-making procedure has the property of being meta-rational.

To appreciate the plausibility of the strong interpretation, consider its negation: that decision-makers adopt decision-making procedures which do *not* maximize meta-utility. If the negative claim were true, then that would mean for at least some decisions, individuals adopt behaviors for which the aggregate decision costs—inclusive of the global decision-making procedure's decision cost—are greater than the aggregate decision costs of some alternative which returns a greater meta-utility value.

But now consider *why* the decision-maker would choose some subordinate decision-making procedure, for which a superior (with respect to meta-utility) alternative existed. Suppose the representative decision-making procedure of a decision-maker were f_n. And suppose that the meta-rational procedure in the representative model $M \in \mathfrak{M}$ were some $f_m'^{M*}$, such that $f \neq f'$. Let $S_0 = f \setminus f'$, and let $S_1 = f' \setminus f$. It must be the case that $S_0 \neq \varnothing \neq S_1$, because $f \neq f'$.

Now consider why a decision-maker would use a decision-making procedure in S_0 rather than a corresponding procedure in S_1. By assumption, the aggregate meta-utilities of the decision-making procedures in S_1 are greater than the aggregate meta-utilities of the procedures in S_0. Surely, the decision-maker would rather use the choices in S_1 if he were only *aware* that the choices in S_1 were superior to the choices in S_0. After all, "utilities" are by definition measures of what the decision-maker seeks to experience; and "decision costs" are measures of what the decision-maker seeks to avoid. Thus is it plausible that the reason why he employs the choices in S_0 is that he is *unaware* that superior choices exist. He must be adopting f_n, because the choices in S_1 are inaccessible to him. Of course, they are not absolutely inaccessible, for a helpful advisor might draw the decision-maker's attention to the existence of better decision-making procedures (thereby drastically reducing the decision cost of selecting them from a higher order of decision-making). They are inaccessible in the same sense that an A. Lange und Söhne chronograph is inaccessible to the average person: their "prices" are prohibitively high.

In other words, *discovery* of the better choices in S_1 imposes additional decision costs—the decision cost assigned to $f_m'^{M*}$—such that f_n generates less aggregate decision cost than $f_m'^{M*}$. And the only principled answer to the question *how much less* is that the difference should be sufficient to offset the reduction in meta-utility. But if this were the case, then f_n would *be* the meta-rational procedure—not f_m'—

contradicting the assumption that f_n is not the meta-rational procedure.

Whether the strong or weak interpretation should be accepted is a philosophical question. I am inclined to believe the strong interpretation is plausible. Yet at a minimum, under the meta-rational conception, the weak interpretation is assuredly true; and the strong interpretation is at least more plausible than under conventional conceptions of rationality.

2.3.4 Extensions

Suppose the strong interpretation of the meta-rational actor hypothesis were true. The strong interpretation takes decision-making procedures, utilities, and decision costs as real phenomena. If they are real, then it follows that for every decision-maker, there exists some meta-rational decision-making procedure which is representative of that decision-maker. It is natural therefore to wonder *what* the "true model" of an individual might be.

The meta-model is too underdetermining to answer this question. Theorem 2.2.5 holds merely that some meta-rational procedure in some model will be *extensionally equivalent* to the "true" model. It is an existence theorem. However, to operationalizing the concept requires more than this—it requires a specification for identifying what the "true" model might be.

The concern is more than merely academic. Theorem 2.2.5 assures logical coherence, but it is not predictive. It is little comfort to the scientist, seeking to make predictions, that a predictive model *must exist* among a panoply of predictive models. For any n observations of an individual's behavior, the scientist desires a model, which reliably predicts what the $(n+1)^{\text{th}}$ observation would be. Theorem 2.2.5 guarantees that whatever the $(n+1)^{\text{th}}$ observation turns out to be, there exists a meta-rational model which *would have* predicted it. But this only pushes the problem back a step, for the question simply becomes whether it is possible to predict which model will predict behavior. To say that *some* prediction is true in any state of the world only begs the question: how to we predict which prediction will be the true one?

As formulated in this article, the meta-model cannot determinatively answer this question. However, it can be a helpful framework upon which to build a satisfactory answer. Specifying *which* among the infinitude of behaviorally consistent models will be predictive (i.e., which determines what the $(n+1)^{\text{th}}$ observation will be) is clearly a valuable scientific objective. If the strong interpretation of the rational actor hypothesis under the meta-rational conception is true, then this imposes one constraint on the potentially valid representations. That the representative decision-

making procedure is meta-rational in the representative model. Yet this is still underdetermining. More criteria are needed.

Some plausible theoretical extensions to the meta-model are: (1) some constraint on the range of decision cost assignments; (2) some constraint on the total number of decision-making procedures chosen; and (3) some constraint on total decision costs. Developing a rigorous extension along these lines would be a substantial undertaking, which would greatly exceed the intended scope of this article. I offer only a sketch of how such an extension might be realized.

The first constraint—on the range of decision cost assignments—reflects the intuition that decision costs are unlikely to vary wildly when one decision-making procedure is used over a domain of many similar states of the world. For example, *whatever* decision-making procedure is used to choose whether to take the elevator or the stairs on one day, that procedure is *ceteris paribus* unlikely to be vastly more or less costly to employ when faced with a similar decision on some other day.

This of course requires a formal definition of the "similarity" relation. A possible first step is to treat conditions not as primitives, but rather as sets of fact-claiming propositions. By "fact-claiming" proposition, I mean to exclude tautological claims like true mathematical propositions and true logical propositions. Let $p_0, p_1, \ldots$ represent fact-claiming propositions, and let $\mathfrak{P} = \{p_0, p_1, \ldots\}$ be the set of all fact-claiming propositions. Now, "conditions" or "states of the world" may be defined as any set of propositions such that $\forall p_i, p_j \in \mathfrak{P}\,((p_i \iff \neg p_j) \implies (p_i \in c \vee p_j \in c))$. A "triggering condition" for any decision-making procedure $f_i \in \mathfrak{L}$ may be defined as $\bigcap Dom(f)$.[43]

The similarity relation κ may now be defined so that $\ulcorner \kappa(c_i, c_j, c_k) \urcorner$ is true iff $|c_i \cap c_j| > |c_i \cap c_k|$. In words, c_i is more similar to c_j than c_k iff the set of propositions which are true of both c_i and c_j has a greater cardinality than the set of propositions which are true of both c_i and c_k.

The second constraint—on the total number of decision-making procedures implicated by the representative global procedure—is motivated by a desire to avoid *ad hoc* decision cost assignments, such that every decision is decided by a unique decision-making procedure. It seems obvious that decision-makers do not have in their repertoire of decision-making procedures a unique decision-making procedure for every state of the world. Decision-making procedures tend to be general and

[43]It is worth noting that the extension which describes conditions as composed of propositions allows the hierarchy to represent logical conditionals. Altering the domain of subordinate decision-making procedures allows for expression of logical negation and logical disjunction. This is sufficient to support a propositional logic of decision-making. The hierarchy may thus be interpreted as a set of logical propositions, prescribing actions given a set of contingent propositions.

applicable over a range of possible conditions.

There are a number of ways this constraint could be modeled. For example, consider two decision-making procedures f_i and f'_j such that $Dom(f) = Dom(f')$. Assume the rank of both decision-making procedures is greater than 1. The set of subordinate decision-making procedures in f_i is given by $Img(f)$, and the set of subordinate decision-making procedures in f'_j is given by $Img(f')$. The constraint may be expressed as a general tendency that if $Img(f) > Img(f')$ then $\forall c\,(d(c, f_i) > d(c, f_i))$. In words, a higher order decision-making procedure which boots to a larger number of subordinate procedures will tend to have a higher decision cost.

The third constraint—that total decision costs should be minimized—is based upon the intuition that choices depend to some extent upon the conditions under which they are decided. Very high decision costs have the capacity to overwhelm the expected utility term in a meta-utility function, rendering it effectively otiose. However, decision-makers do seem in very many circumstances to be sensitive to the expected utility of their choices. Therefore, for a given decision-making procedure f_i, the aggregate decision costs assigned, i.e., $\sum_{c \in Dom(f)} (\mu(c, f_i, M) - EU(c, \theta(c, f_i)))$, should not be greater than necessary.

Of course, it is clearly the case that decision costs *can* utterly overwhelm the expected utility term in a meta-utility function. To illustrate, consider the cryptocurrency Bitcoin. The value of a Bitcoin token is approximately $6,500 as of November, 2018. The Bitcoin address holding the greatest number of tokens is "3D2oetdNuZUqQHPJmcMDDHYoqkyNVsFk9r,"[44] which has a balance of approximately 140,000BTC, for a total fiat value of more than $900 million. Any individual could, of course, simply transfer all the tokens to himself. And it is plausible that individuals would tend derive quite a lot of utility from possessing $900 million. All the thief would need is the "private key" associated with the address. However, discovery of the private key is—in the language of complexity theory—an *NP*-Complete problem. The decision cost is astronomically high. So high, in fact, that it is not worth anyone's effort to attempt the task, despite the expected payoff of $900 million. And increasing the payoffs tenfold, a hundredfold, or a thousandfold are unlikely to alter behavior, because the decision costs associated with hacking private keys are so great that any meta-rational procedure will tend to select some "Don't bother hacking Bitcoin" heuristic. This, indeed, is the very point of cryptocurrencies. They rely upon decision costs—rather than armed guards—to protect holdings.

Nevertheless, it is reasonable to suppose that few of the decisions that real individuals routinely face are comparable to hacking Bitcoin private keys. In most cases, increasing or decreasing the expected payoffs of outcomes will tend to influence the

[44] The address is the "cold wallet" of the cryptocurrency exchange, BitFinex.

choices they would select.

The three constraints, taken together, could be used to determine which meta-rational models are more or less likely to be predictive. One possible mechanism for operationalizing the three proposed constraints would be to formulate a "fit" function, measuring the extent to which a given model violated the three constraints. More specifically, the factors would be: (i) the variance of decision costs assigned to a decision-making procedure, (ii) the number of subordinate decision-making procedures, and (iii) the total aggregate decision costs implicated by a representative decision-making procedure. The fit function would return, for every model and meta-rational procedure within that model, some quantitate value, representing the degree of "fit" within the constraints. Increasing variance in the assignment of the decision costs of a decision-making procedure over triggering conditions would decrease a model's "fit" value. Decreasing decision costs assigned to a decision-making procedures, which rely on a greater number of subordinate procedures would decrease a model's "fit" value. And increasing total decision costs implicated by a meta-rational procedure would decrease a model's "fit" value. Of course, weights could be added to emphasize the relative importance of the three constraints.

The maxima of the "fit" function would determine the models which are most likely to be predictive. Such an approach may usefully operationalize the meta-rational conception, identifying those models described by the meta-model, which are most likely to be representative of real decision-making. Minimizing the role that decision costs play in the formulation of meta-utility is therefore a facially plausible theoretical presumption.

Of course, the most forcing constraints would require an empirical investigation of actual human behavior, charting out the terrain of decision-making procedures and decision costs. There is a limit to how far we can reason from first principles. At some point, predicting behavior requires an investigation into how real people behave in the real world.

It is important to observe that further extensions to the meta-model would not require a program of research vastly different from what behavioral economists and psychologists already do. The identification of common behavioral patterns—in particular non-rational behavioral patterns—is precisely the kind of terrain-mapping which would lead to a more precise characterization of decision-making procedures and decision costs. The meta-rational approach is not revolutionary. It merely imposes a theoretical structure on the work of behavioral researchers, resolving some of the conceptual problems latent in the underlying assumptions.

Yet this article's contribution is not wholly impotent, for it indicates some gaps in the present state of the research. While the volume of research identifying various

decision-making procedures continues to grow steadily, little attention has been given to locating the ranks of decision-making procedures in the hierarchy. Neither has there been much effort directed at measuring decision costs, identifying the tipping points at which rational deliberation becomes too costly, and alternative decision-making procedures are chosen.

Concededly, "tipping point" tests may often be difficult to perform in the lab. In principle, there are two straightforward mechanisms for measuring decision costs. The first involves a manipulation of decision costs, posing approximately analogous decision problems with equivalent payoffs, increasing in complexity. Meta-rationality suggests that as the difficulty of the problem is increased, subjects will tend to abandon rational deliberation in favor of some less costly decision-making procedure. The payoffs of the decision problem can then be used to estimate a disutility curve increasing in problem complexity, representing the decision cost assignment of rational deliberation. The process can then be continued to induce further changes from fine-grained heuristics to more coarse-grained heuristics, progressively sacrificing a greater proportion of expected payoff for reductions in decision costs.

The basic strategy is attractive when stated in vague terms. However, it is difficult to imagine how a rigorous experiment might be devised, which progressively increases the decision cost of rational deliberation. Many assumptions would be required, and designing a satisfactory experimental setup would be a non-trivial problem.

The second approach to determining a "tipping point" would be to vary payoffs to observe changes in decision-making procedures. The idea is that when decision-makers are faced with outcomes, which diverge dramatically in payoffs, avoiding the cost of choosing an inferior action will tend to justify a greater investment in decision-making. Thus, increasing payoffs should, ceteris paribus, tend to cause individuals to behave more rationally.

There have been few studies, pursuing the second strategy, presumably because testing meaningfully increasing stakes would require more funding than experiments with adequate sample sizes are able to access. The evidence thus far uncovered has been inconclusive,[45] and it is certainly insufficient for the purposes of quantifying decision costs. Nevertheless, it is difficult to overestimate the inventiveness of experimentalists, and it remains entirely possible that less obvious mechanisms for measuring decision costs may yet be devised.

It is worth remarking that meta-rational models are not *only* descriptive. A meta-rational procedure also *prescribes* how decision-makers *should* behave in the presence of decision costs. In contexts where the relevant functions are known, the

[45] *See, e.g.*, Diekmann, 2004, Smith and Walker, 1993, Fehr-Duda, Bruhin, Epper, and Schubert, 2010.

normative implications can be extremely useful. Three obvious contexts in which the normative analysis could prove fruitful are: the analysis of legal decision-making, operations research, and machine decision-making.

Where a model is *given*, and if there exist meta-utility suprema in $\mathfrak{L}$, the identification of meta-rational procedures will determine how decision-makers ought to behave. In the law, decision costs are helpfully monetized in the form of court costs. The robust literature in the Law & Economics literature studying the second order decision whether to employ "rules versus standards" analogizes well to the meta-model.[46] Further exploration, for example, might investigate what third order rule should be used to determine what second order rule should be used to determine whether a "rule" or a "standard" should be used.

However, a normative operationalization of meta-rational models would likely require substantial simplification in order to make it tractable. These simplifications could be implicit in the models given. For example, in the context of legal decision-making, there will exist minimum fixed costs of an adjudicative hearing. These minimum fixed costs would effectively truncate the infinitely large hierarchy of decisions to some finite maximum order, above which higher orders of decision-making could be ignored.

2.4 Conclusion

The meta-rational meta-model resolves several critical conceptual problems undermining the rational actor hypothesis. In particular, the infinite regress problem, idempotent hierarchy problem, explanatory vacuousness of generalized expected utility theories, and predictive fragility of generalized expected utility theories. In addition, it provides foundation to conventional expected utility analysis, second-order models of bounded rationality, and psychological researches in heuristics.

The meta-model also provides theoretical support for the directions which behavioral researches have pursued over the past several decades, while suggesting refinements for future research, such as an increased focus on the study of hierarchical order and decision costs.

The principal result is to establish that any set of potential behaviors can be represented in some model as a meta-rational procedure. Individuals are necessarily representable as "rational" actors. It suggests moreover the plausibility of the stronger hypothesis: that individuals are not only *representable* as meta-rational, but that they *are* in fact meta-rational.

[46] For example, see Kaplow, 1992.

Bibliography

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, *21*(4), 503–546.

Alt, F. (1971 [1936]). On the measurability of utility. In *Preference, utility and demand* (pp. 424–431). Harcourt Brace.

Baron, J. (2000). *Thinking and deciding* (3rd Ed.). Cambridge University Press.

Becker, G. (1976). *The economic approach to human behavior*. University of Chicago Press.

Birnbaum, M. H. (2004). Test of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: Effects of format, event framing, and branch splitting. *Organizational Behavior and Human Decision Processes*, *95*, 40–65.

Birnbaum, M. H. (2006). Evidence against prospect theories in gambles with positive, negative, and mixed consequences. *Journal of Economic Psychology*, *27*, 737–761.

Birnbaum, M. H. (2008a). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463–501.

Birnbaum, M. H. (2008b). New tests of cumulative prospect theory and the priority heuristic: Probability-outcome tradeoff with branch splitting. *Judgment and Decision Making*, *3*(4), 304–316.

Birnbaum, M. H., & Navarrete, J. B. (1998). Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty*, *17*, 49–78.

Camerer, C., & Thaler, R. H. (1995). Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives*, *9*(2), 209–219.

Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature*, *34*(2), 669–700.

Diekmann, A. (2004). The power of reciprocity: Fairness, reciprocity, and stakes in variants of the dictator game. *Journal of Conflict Resolution*, *48*(4), 487–505.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, *75*(4), 643–669.

Fehr-Duda, H., Bruhin, A., Epper, T., & Schubert, R. (2010). Rationality on the rise: Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty*, *40*(2), 147–180.

Field, H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy*, *70*(14), 462–481.

Gigerenzer, G. (2001). The adaptive toolbox. In *Bounded rationality*. MIT Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Hirshleifer, J. (1985). The expanding domain of economics. *American Economic Review*, *75*(6), 53–68.

Jolls, C., Sunstein, C. R., & Thaler, R. H. (1998). A behavioral approach to law and economics. *Stanford Law Review*, *50*, 1471–1550.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193–206.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kaplow, L. (1992). Rules versus standards: An economic analysis. *Duke Law Journal*, *42*, 557–629.

Korobkin, R., & Ulen, T. S. (2000). Law and behavioral science: Removing the rationality assumption from law and economics. *California Law Review*, *88*(4), 1051–1144.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lazear, E. P. (2000). Economic imperialism. *Quarterly Journal of Economics*, *115*(1), 99–146.

List, J. A. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica*, *72*(2), 615–625.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. Oxford University Press.

Minsky, M. (1986). *Society of the mind*. Simon and Schuster.

Al-Najjar, N. I., & Weinstein, J. (2009). The ambiguity aversion literature: A critical assessment. *Economics and Philosophy*, *25*, 249–284.

Plott, C. R., & Zeiler, K. (2007). Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory? *American Economic Review*, *97*(4), 1449–1466.

Savage, L. J. (1954). *The foundations of statistics*. John Wiley & Sons.

Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138.

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, *31*(2), 245–261.

Sunstein, C. R., & Ullmann-Margalit, E. (1999). Second-order decisions. *Ethics*, *110*, 5–31.

Suppes, P., & Winet, M. (1955). An axiomatization of utility based on the notion of utility differences. *Management Science*, *1*, 259–270.

Thaler, R. H. (2015). *Misbehaving: The making of behavioral economics*. W. W. Norton & Company.

Triantaphyllou, E., & Mann, S. H. (1989). An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox. *International Journal of Decision Support Systems*, *5*(3), 303–312.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton University Press.

Wright, J. D., & Ginsburg, D. H. (2012). Behavioral law and economics: Its origins, fatal flaws, and implications for liberty. *Northwestern University Law Review*, *106*(3), 1033–1090.

Zamir, E. (2018). Reinforcing law and economics: Behavioral support for the predictions of standard economic analysis. *Hebrew University of Jerusalem Legal Research Paper No. 18–17*.

Chapter 3

Bounded Criminality

The vast majority of the Behavioral Law & Economics literature focuses on the mitigation of cognitive biases and heuristics. Indeed, "debiasing" and "insulation" are practically assumed to be sound policy goals, *ipso facto*, wherever biases and heuristics may be found. Of course, no one disputes that self-interested rationality is a valid *normative* model. The "revolutionary" claim of behavioral economics is not that the rational actor hypothesis fails to describe how people *ought* to behave, but rather that the rational actor hypothesis fails to describe how people *do* behave in fact. Thus, the thinking goes, strategies to eliminate non-rational behavior will close the gap between how people ought to act and how they do in fact act.

Undoubtedly, debiasing and insulating strategies do tend to effect efficiency, assuming perfect competition. However, the vast majority of markets routinely suffer from one or more forms of market failure—negativing the assumption of perfect competition. In such cases, we may want, as social engineers, to increase the gap between privately optimal behavior and behaviors in fact. Where the private objective and social objective fail to align, it may be possible to exploit systematic non-rational behavior to effect second order incentive alignment. That is, to use predictable and systematic non-rational behavior as a tool to correct for market inefficiencies. Lamentably (and perplexingly), few scholars have yet seized upon the exploitation of non-rational behavior as a method of effecting incentive alignment.

In this paper, I investigate this possibility in the context of criminal law policy, arguing that superior deterrent effects may be achieved with less enforcement and more lenient sentencing by exploiting the phenomenon of "bounded rationality." Section 3.1 briefly summarizes the history of economic theories of criminal law, and also addresses some historical objections to the premises underlying economic approaches to criminal law. Section 3.2 describes how the conventional economic conception of

criminal deterrence may be extended to incorporate systematic non-rational behavior. Section 3.3 cashes out that conceptual work in a formal model, from which an alternative "efficient" level of criminal deterrence may be computed. In Section 3.4, I suggest some possible practical applications of my model in policymaking. Section 3.5 concludes with an overview of what I have tried to accomplish in this essay.

3.1 Background

3.1.1 The Elements of Criminal Deterrence

The Enlightenment era legal philosopher Cesare Beccaria was among the first instrumentalists, arguing that punishments should not exceed the net benefits derived from incapacitation and deterrence.[1] However, the importance of Beccaria's work extends well beyond his philosophical rejection of retribution as a ground for punishment. His analysis of deterrence implicitly relied on economic methods and results, an approach which would reemerge two hundred years later when scholars in the Law & Economics movement began systematically applying economic methods to the study of criminal law. Yet Beccaria's insights remain a pellucid description of the fundamental framework Law & Economics scholars employ in analyzing the criminal law; the wealth of insights bears substantial direct quotation:

> It is not only the common interest of mankind that crimes should not be committed, but that crimes of every kind should be less frequent, in proportion to the evil they produce to society. Therefore the means made use of by the legislature to prevent crimes should be more powerful in proportion as they are destructive of the public safety and happiness, and as the inducements to commit them are stronger. Therefore there ought to be a fixed proportion between crimes and punishments.
>
> It is impossible to prevent entirely all the disorders which the passions of mankind cause in society. These disorders increase in proportion to the number of people and the opposition of private interests. If we consult history, we shall find them increasing, in every state, with the extent

[1] Cesare Beccaria, Dei delitti e delle pene [Of Crimes and Punishments], Chapter 12, Edward D. Ingraham, trans. (1764) [1819] ("The end of punishment, therefore, is no other than to prevent the criminal from doing further injury to society, and to prevent others from committing the like offence. Such punishments, therefore, and such a mode of inflicting them, ought to be chosen, as will make the strongest and most lasting impressions on the minds of others, with the least torment to the body of the criminal.").

> of dominion. In political arithmetic, it is necessary to substitute a calculation of probabilities to mathematical exactness. That force which continually impels us to our own private interest, like gravity, acts incessantly, unless it meets with an obstacle to oppose it. The effect of this force are the confused series of human actions. Punishments, which I would call political obstacles, prevent the fatal effects of private interest, without destroying the impelling cause, which is that sensibility inseparable from man. The legislator acts, in this case, like a skilful architect, who endeavours to counteract the force of gravity by combining the circumstances which may contribute to the strength of his edifice.[2]

Given the private welfare function, $P(a_n) = B(a_n) - K(a_n)$, where $B(a_n)$ is the benefit of undertaking action a_n, and $K(a_n)$ is the the cost of undertaking a_n, it may arise that the optimal choice, $a^* = \max_{a_n} P(a_n)$ is a criminal act. This is, in Beccaria's words, the "gravity" that impels people to commit crime: it is simply the choice that maximizes their payoffs. However, by imposing a criminal sanction, $S(a_n)$, discounted by the probability of enforcement π, the law manipulates parties' cost-benefit calculations by imposing "political obstacles," such that:[3]

$$P^{\dagger}(a_n) = B(a_n) - K(a_n) - \pi(a_n)S(a_n) \tag{3.1}$$

The idea is that when the expected sanction $\pi(a_n)S(a_n)$ is sufficiently large, then $a^* \neq a_n$, and parties will voluntarily refrain from undertaking the proscribed activity a_n, since it will no longer be privately optimal.

3.1.2 Criminological Skepticism about Economic Models of Deterrence

Before proceeding to modern approaches to criminal deterrence, I will consider and rebut some objections to Beccaria's rudimentary formulation in the recent criminal law literature.[4] I have included this discussion with some hesitation, and it bears remarking on what role this plays in my argument. Readers of early drafts

[2] Beccaria, *supra* note 1, Chapter 6.

[3] Observe that in an alternative framing of the problem, $K(a_n)$ could be interpreted to include $\pi(a_n)S(a_n)$. I have chosen to treat the terms separately for analytical clarity, though prior Law & Economics scholars writing on criminal law have typically considered $\pi S(a_n)$ to be an element of $K(a_n)$. The choice to treat the two terms separately is motivated merely by my preference for clarity over parsimony, and obviously has no effect on the results.

[4] It bears observing (if only because there yet remains a dwindling but still sizable number of lawyers with little or no exposure to Law & Economics) that numerically, the overwhelming

of this paper have given vastly differing opinions on the importance of addressing this point. Those whose backgrounds were in Law & Economics complained that it was an unnecessary digression, while those whose backgrounds were in criminology complained that it was too cursory. So much reveals the methodological divide in criminal law scholarship. Upon consideration, it seems to me that some defense of a critical premise—that punishment has a deterrent effect—is warranted, since at least some portion of readers, whom I hope to reach, will regard a defense of the economic approach to criminal deterrence an essential precondition to further theoretical elaborations. Yet it is not my objective to conclusively establish this point here, and an exhaustive investigation of this issue would lead us far astray. Rather, I should like simply to acknowledge the controversy, point to several persuasive arguments against the criminologist's skepticism, and proceed quickly to the novel contributions to the economic approach I propose in Section 3.2.

According to the Beccarian formulation, the expected sanction consists of two terms: the probability of enforcement π, and the magnitude of sanction $S(a_n)$. If the simple model, $P^{\dagger}(a_n) = B(a_n) - K(a_n) - \pi(a_n)S(a_n)$ is correct, then it follows that, *ceteris paribus*, a $1/2$ reduction in the probability of enforcement would be wholly offset by a doubling in the magnitude of sanction, and vice versa. That is, by simple arithmetic:

$$\frac{\pi(a_n)}{2} \times 2S(a_n) = \pi(a_n)S(a_n) = 2\pi(a_n) \times \frac{S(a_n)}{2} \tag{3.2}$$

However, there exists a substantial body of empirical research suggesting that the

majority of "objections" to the economic approach are not of the learned variety described in this subsection. Rather, the predominant strain of critiques of the economic analysis of criminal law (and indeed, of economic analysis generally) seem to arise from the puerile and abysmally ignorant contention that economic models are unsound because "people don't think like that." While such woefully stupid misunderstandings hardly merit reply, I shall in a spirit of charity offer the hopefully edifying reminder that economic models are not meant to provide a psychological account of how people *reason*, but rather an extensional account of how they *behave*. Thus, regardless of whether people consciously perform cost-benefit calculations to determine their actions, or whether they are "subconsciously" impelled toward utility-maximizing choices by instinct or habit is immaterial from an economic perspective. That the "failure" of economics to track psychology somehow represents a defect in the theory is based on a gross misconception.

It is true, behavioral economics poses a challenge to rational choice economics, and behavioral economics is founded on "realistic" psychology rather than the Herculean ideal of *homo economicus*. However, the conflict between rational choice theory and behavioral economics does not arise because behavioral economics furnishes a better account of how people *think*, but rather because it seems to provide a better prediction of how people will *act*. Never and nowhere is economics concerned with what goes on in people's heads, except insofar as it provides convenient indices to how they will ultimately *behave*.

equivalencies in Formula 3.2 fail to obtain in the real world.[5] Instead, the evidence indicates that increasing the probability of enforcement, π, is vastly more effective than ratcheting up the magnitude of prescribed punishment, $S(a_n)$,[6] implying:[7]

$$\frac{\pi(a_n)}{2} \times 2S(a_n) < \pi(a_n)S(a_n) < 2\pi(a_n) \times \frac{S(a_n)}{2} \tag{3.3}$$

It is first worth pointing out that, even if this were generally the case, it would not necessarily be as "fatal" for a severity-as-deterrence approach to criminal punishment as some critics might suppose. It would merely require a trivial modification of the sanction function—assuming the sociological data were sufficiently strong to warrant a modification of the formulation.[8]

However, it would be overhasty to infer Formula 3.3 from the raw sociological data. Despite the eagerness of some criminal law scholars in declaring the "death" of severity-as-deterrence and the economic analysis of criminal law generally,[9] to reject economic methodology because it fails to perfectly predict real-world effects would be as silly as rejecting Newton's laws of motion simply because they fail to predict the travel of a struck tennis ball. While it is true that spin, air pressure, lift, and friction generally will create complications to the extent that the precise flight path of

[5] *See, e.g.,* Andrew von Hirsch, et al., Criminal Deterrence and Sentence Severity: An Analysis of Recent Research (Hart Publishing, 1999); Paul Langan & David Farrington, "Crime and Justice in the United States and England and Wales, 1981-96," Bureau of Justice Statistics (1998); David Farrington, Paul Langan and Per-Olof H. Wikstrom, "Changes in Crime and Punishment in America, England and Sweden between the 1980s and the 1990s," 3 Studies in Crime Prevention 104-131 (1994); Paul Gendreau, Claire Goggin & Francis Cullen, "The Effects of Prison Sentences on Recidivism," Ottawa, Ontario, Canada: Public Works and Gov't Services Canada (1999) (finding that increasing prison sentences actually correlated with a 3% *increase* in recidivism). *See generally* Cheryl Marie Webster & Anthony N. Doob. "Searching for Sasquatch: Deterrence of Crime Through Sentence Severity" in Oxford Handbook of Sentencing and Corrections 191, 173-195 (Joan Petersilia & Kevin R. Reitz, eds. 2012) (surveying the empirical research contradicting deterrence theories' predictions).

[6] *See generally* Michael Tonry, *Learning from the Limitations of Deterrence Research*, 37 Crime and Justice 279-311 (2008).

[7] Or, more generally, $\frac{\pi(a_n)}{x} \times xS(a_n) < \pi(a_n)S(a_n) < x\pi(a_n) \times \frac{S(a_n)}{x}$.

[8] For example, $D(S) = \alpha S - \beta S^2$, such that increasing criminal penalties $S(a_n)$ would yield diminishing effects (i.e., $\frac{\partial D}{\partial S^2} = -2\beta$).

[9] *See, e.g.,* Tonry, *supra* note 6, at 280 ("[M]acro-level modeling of deterrent effects of changes in sanctions policies by economists and econometricians has reached a dead end, as Ronald Coase in 1978 predicted would happen concerning subjects on which the economist's advantage was primarily one of technique."); Cheryl Marie Webster & Anthony N. Doob, *supra* note 5, at 191 ("[T]he continued centrality of this deterrence theory as a sentencing objective constitutes a false promise, contributing to a waste of resources and a reduction in the public's confidence in the criminal justice system, while encouraging policy makers to ignore more effective crime control strategies.").

a tennis ball will be practically incalculable, the principles of classical mechanics are hardly "refuted" thereby; the laws of physics reveal a great deal about nature, even when they fail at prediction. Likewise, the results of an economic theory should not be rejected simply because they fail to account for every force operating in the real world. The mere fact that empirical results diverge from theoretical predictions may only signal that additional forces are at work, and that noise from these other forces prevents us from isolating the processes-of-interest in our observations, in which case the proper approach would not be to abandon what progress has been made, but rather to identify and describe the other factors in play.

Lamentably, a significant contingent of criminal law scholars have quite exuberantly declined to pursue this eminently reasonable approach, opting instead to accuse economists of ideological bias and fallacious reasoning. For example, Michael Tonry writes, "[S]ome or much of the work on deterrence by economists may be conscious or unconscious products of ideological, as opposed to merely disciplinary, ways of thinking. . . . Many of the economists who have written on the deterrent effects of punishments are well-known political conservatives—Gary Becker, Richard Posner, Isaac Ehrlich, John Lott—and others such as Joanna Shepherd are less well-known conservatives," although evidently in a magnanimous mood, he adds, "It is merely human to be deeply attached to one's intuitions."[10]

In addition to such overeager announcements of the demise of economic theories of criminal law, such criticisms abound with fundamental misunderstandings about economic modeling. For example, Cheryl Marie Webster and Anthony Doob write, "[Deterrence through severity] strategies assume that potential offenders conduct sophisticated analyses of the relative costs of various penalties."[11] This is hardly a new criticism of the rational actor hypothesis, though it is easily answered. The rational actor hypothesis does not assume that people consciously conduct sophisticated cost-benefit calculations. The model is not intended to track deliberation, but rather *behavior*. The reason why people, on average, will tend to behave in a manner consistent with the rational actor model is not because they are consciously deliberating about welfare maximization, but rather because the net effect of gut intuitions, genetic predispositions, and environmental conditioning ultimately cash out in choices and actions which coincide with payoff maximization. A particular actor may articulate reasons for her actions very different from welfare optimization—and we need not say that one is a "right" or "wrong" account of the decision-making, except insofar as it succeeds or fails at explaining behavior. Certainly, it is problematic for the rational actor model if its predictions fail to track real-world decisions—as

[10]Tonry, *supra* note 6, at 304-305.
[11]Webster & Doob, *supra* note 5, at 182.

appears to be the case (at least superficially) with severity-as-deterrence—but it is not problematic that potential offenders are not *consciously* performing cost-benefit analyses, as Webster and Doob complain.

Happily, it turns out that a number of the "frictions" at work are easily identifiable. First, the portion of the population that will commit at least one felony in their lives is small—and they may simply be dismissed as outliers. This is not the most satisfying explanation for such deviant behavior, but neither is it unwarranted. It may simply be that criminals have highly idiosyncratic risk-preferences.[12] Another explanation may be time-inconsistent discounting.[13] Yet another possible explanation may be that criminals are psychologically *incapable* of rationally responding to incentives.[14]

A different cluster of rationales considers whether it might simply be that particular social circumstances make criminal activity optimal for some people,[15] even when the cost of expected sanctions is rationally calculated—in which cases, incidentally, it is likely that the collateral effects of conviction will exacerbate the probability of recidivism, since the effects of post-incarceration collateral consequences tend to

[12] *See, e.g.,* Cathy Buchanan & Peter R. Hartley, *Criminal Choice: An Economic View of Life Outside the Law*, Policy 54-58, *Autumn* (1990) ("Results from studies in expected utility theory suggest that the more risk-averse the individual is, the less he will like an increase in penalties compensated by a reduction in capture probability. . . . On the other hand, risk-loving individuals will be deterred more by increases in the probability of capture than compensating increases in penalties. Since crime is a risky occupation, risk-loving and less risk-averse individuals will find it a more satisfactory employment.").

[13] *See, e.g.,* Manuel A. Utset, *When Good People Do Bad Things: Time-Inconsistent Misconduct & Criminal Law*, FSU Public Law Research Paper No. 232 (2006).

[14] *Id.* at 57 ("Some crimes are committed by people who are either temporarily or permanently insane, acting in a fit of passion or under the influence of drugs."); Sheilagh Hodgkins, *Mental Disorder, Intellectual Deficiency, and Crime Evidence from a Birth Cohoert*, 49 Arch Gen Psychiatry 476-483 (1992) (finding that men suffering from major mental disorders were four times more likely to be registered for violent offenses, while women with major disorders were 27 times more likely to be violent offenders—though statistically, this still only accounts for a small minority of prison populations).

[15] John R. Lott, Jr., *A Transaction-Costs Explanation for Why the Poor Are More Likely to Commit Crime,* 19 J. Legal Stud. 243 (1990), observes that because bankruptcy and antislavery laws impose a transaction cost on lending to individuals, whose primary asset is bare human capital (i.e., since debtors cannot be enslaved to extract repayment through forced labor, and also because bankruptcy shields them from debt recovery in certain cases, banks face an increase in the probability that loans will go unpaid), it may be that for some persons the transaction cost of theft may be sufficiently below the transaction cost of borrowing, that theft will remain privately optimal even when penalties are optimal. Ironically, the tradeoff then becomes that prohibiting debt-recovery-through-slavery gets replaced by increased rates of incarceration, and if appears that the most economically disadvantaged end up in chains either way!

decrease the opportunity costs of crime.[16]

Another source of "friction" is that increasing severity of sanctions fails to affect parties' private welfare calculations when they are unaware of what the law is—which is likely the case for the vast majority of the population.[17] It is important to keep in mind that the function π represents the *perceived* probability of enforcement, and $S(a_n)$ represents the sanction *believed* to follow from a_n. Where the legal consequences of lawbreaking are insufficiently publicized, it is no defect of the Becarrian model (or a superficial defect at worst) that deterrence effects fail to obtain. It is trivially the case that actors will not be responsive to imperceptible incentives.[18]

Ultimately, although the point is not utterly without controversy,[19] even if we take it as given that the empirical data does not support severity-as-deterrence predictions in society as it presently is, this need not necessarily be interpreted as a refutation of the theory. More than 90% of the U.S. population will never commit a felony in their lives,.[20] and it may simply be the case that a severity-as-deterrence model

[16] Alec Ewald & Christopher Uggen, *The Collateral Effects of Imprisonment on Prisoners, Their Families, and Communities,* Chapter 3 in Oxford Handbook of Sentencing and Corrections (Joan Petersilia & Kevin Reitz, eds., 2012). *See also* John Schmitt & Kris Warner, *Ex-offender and the Labor Market*, Center for Economic and Policy Research (November 2010).

[17] *See* Kirk R. Williams, Jack P. Gibbs, and Maynard L. Erickson, *Public Knowledge of Statutory Penalties: The Extent and Basis of Accurate Perception,* 23 Pacific Sociological Review 105-128 (1980).

[18] I do not mean to suggest that greater public education will *necessarily* result in more effective severity-as-deterrence, though no doubt it could not hurt. Rather, I mean simply to point out another possible explanation for the failure of severity-as-deterrence to manifest in empirical data.

[19] *See, e.g.,* Daniel Kessler & Steven D. Levitt, *Using Sentence Enhancements to Distinguish between Deterrence and Incapacitation*, 42 J. Law & Economics 343-364 (1999) (arguing that California's Proposition 8 demonstrates both the observable effect of deterrence and incapitation in crime rates). *But see* Cheryl Marie Webster, Anthony Doob & Franklin Zimring, *Proposition 8 and Crime Rates in California: The Case of the Disappearing Deterrent*, 5 Criminology & Public Policy 417-448 (2006) (arguing that a more fine-grained analysis reveals that the crime rates began dropping prior to the passage of Proposition 8, and pointing out other methodological problems in Kessler & Levitt (1999)). *But see* Steven D. Levitt, *The Case of the Critics Who Missed the Point: A Reply to Webster et al.*, 5 Criminology & Public Policy 449-460 (2006) (arguing that the data from Kessler & Levitt (1999) withstands the criticisms of Webster, et al. (2006)).

[20] Sarah Shannon, et al., *Growth in U.S. Ex-Felon and Ex-Prisoner Population, 1948 to 2010*, Paper presented at the 2011 Annual Meetings of the Population Assoc. of America (2011) ("By our estimates, about 3.4 percent of the adult voting age population have once served or are currently serving time in a state or federal prison. If we adopt a more inclusive definition of the criminal class, including all convicted of a felony regardless of imprisonment, these numbers increase to 19.8 million persons, representing 8.6 percent of the adult population and approximately one-third of the African American adult male population.").

is only effective for approximately 90% of the population. As I have discussed in the foregoing paragraphs, it seems plausible that the portion of the population who commit felonies may for various reasons be unresponsive to the incentive effects of severity-as-deterrence, and it may be that increased sentence severity does not translate to decreased crime rates for the simple reason that the deterrence effect is already maximal in industrialized nations (indeed, I will later argue that sanctions very likely exceed the point at which marginal deterrence yields *de minimis* returns). On this view, a better test of severity-as-deterrence might be to investigate whether the converse proposition holds: that a reduction in the severity of sanctions results in increased criminal activity.[21]

Finally, a discussion concerning skepticism about economic models in the criminological literature would not be complete without some mention of the sociological data *supporting* the notion that increasing severity of sanctions generates a deterrent effect.[22] In what is likely the most thoroughgoing technical defense of sanction-as-deterrence, Silvia Mendes and Michael McDonald tackle the sociological data supposedly "refuting" the effect of severity directly, arguing that the apparently limited effects of sentence severity on deterrence may be attributed to conceptual errors on the part of the statisticians interpreting the data.

> [T]he dubious findings regarding the inconsistent effect of the severity component of deterrence theory are a consequence of theoretical slippage when moving from the verbal theoretical statement to the statistical representation of that statement. Our purpose is to demonstrate that the failure to include any of the deterrence theory components "unbundles the theoretical package." For this reason, we argue that the empirical ambiguity with respect to sentence severity arises because sometimes the empirical formulation of deterrence theory fails to keep the theoretical package intact. In particular, statistical models that isolate the components through the use of separate, additive elements do not account for the expected cost calculation as specified in the theory. Sentence length does not work independently of the probability of arrest and conviction. Rather, all three elements operate in combination.[23]

[21] Although, I will later argue that the reduction will have to be substantial to detect such effects.

[22] *See, e.g.,* Charles R. Tittle, Sanctions and Social Deviance—The Question of Deterrence (National Science Foundation 1980); Richard C. Hollinger, *Deterrence in the Workplace: Perceived Certainty, Perceived Severity, and Employee Theft*, 62 Social Forces 398-418 (1983).

[23] Silvia M. Mendes & Michael D. McDonald, *Putting Severity of Punishment Back in the Deterrence Package,* 29 Policy Studies Journal 588-610, 590 (2001). *See also* Silvia M. Mendes,

Nevertheless, the balance of opinion among non-economist criminal law scholars seems to be that punishment is less effective than policing, which if true would require at least a minor modification of the the Beccarian cost-benefit formula. As the foregoing discussion suggests, I do not agree that such modification is warranted, though I will argue for a more nuanced understanding of costs and benefits in Section 3.2 for different reasons.

3.1.3 Becker's Approach to Crime

Beginning in 1968, with the publication of Gary Becker's seminal article, *Crime and Punishment: An Economic Approach*,[24] there has been a steady flow of research on criminal law and punishment from an economic perspective. The main line of inquiry has followed Beccaria[25], Bentham,[26] and Becker,[27] in treating instrumentalist objectives (and deterrence foremost) as the exclusive grounds for criminal punishment.

Becker's model remains the foundation, upon which most subsequent economic theories of criminal law have built. Becker's model treats crime in price theoretic terms. The basic setup of the model begins with the premise that actors seek to maximize their private welfare. In some circumstances, the optimal choice may be a criminal act. In such cases, the actor's choice create an externality (i.e., the harm caused to victims). In the aggregate, the criminal's benefit and the victim's loss, taken together, will usually represent a net loss.[28]

Policing and punishment create disincentives against the commission of crime (*See* Formula 3.1). The effect is analogous to increasing prices on demand. The effect of the law here is to increase the analogical "price" of criminal activity. Naïvely, therefore, our first intuition may be that punishments for all crimes should be maximal. The countervailing consideration however is that policing and punishment are

Certainty, Severity, and Their Relative Deterrent Effects: Questioning the Implications of the Role of Risk in Criminal Deterrence Policy, 32 Policy Studies Journal 59-74 (2004).

[24] 76 J. Pol. Econ. 169 (1968).

[25] Beccaria, *supra* note 1

[26] Jeremy Bentham, "An Introduction to the Principles of Morals and Legislation," in 1 Works of Jeremy Bentham 1, 86-91 (J. Bowring ed. 1843).

[27] Becker, *supra* note 24.

[28] Why this should be the case may not be immediately obvious, since theft, for example, is merely a transfer of wealth; and transfers of wealth are, *ceteris paribus* neutral for social welfare (assuming a Kaldor-Hicks aggregation criterion). One compelling explanation is that unchecked criminal "transfers of wealth" create a rent-seeking scenario, since criminals must expend resources to obtain their loot, while victims will responsively expend resources to guard their property. Since the "prize" remains a fixed quantity, the effect is nonproductive competition. *See generally* Becker, *supra* note 24.

costly.[29] Thus, the price cannot be raised indefinitely, and optimal enforcement (i.e., where "enforcement" consists of policing and punishment) may be determined to be the point, at which the marginal cost of enforcement is equal to the marginal benefit of crime-reduction.

Translated into everyday terms, Becker's claim seems eminently sensible. If the destructiveness of crime costs more than enforcement, society should increase policing and penalties. If, on the other hand, things get to a point where we spend more to prevent crimes than the cost of simply allowing the crime to occur, then we have gone too far. The proper investment in crime-prevention is such that society receives an equal return in the reduction of the cost of crime. So far, so good.

3.2 A New Framework for Deterrence

The Becker model relies upon the "rational actor hypothesis," which posits that on average, people tend to make utility-maximizing choices. This view of human behavior has come under intense attack in recent decades, although skepticism about it has existed since the earliest days of economics.[30] Indeed, it would be a mistake to infer from Becker's assumptions that *he* believes that they are *true*,[31] at least no more than a cartographer believes that Earth is flat.[32] A theory, by necessity, makes simplifying assumptions. This should not be seen as a concession, but rather the very *purpose* of theory-building. Were a map perfectly similar to the thing it was meant to represent, it would cease to be a representation, and would be instead a duplicate. Maps help us to understand things by "flattening" reality and transforming it in ways that preserve some relations in the world, while distorting others, to pick out salient information for navigation. This principle applies to theoretical models no less than to geographical representations.

Nevertheless, while this analogy may quell the most puerile criticisms of the

[29] Punishment is not always costly. For example, fines carry no social cost.

[30] *See, e.g.,* Francis Edgeworth, Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences, 16 (1881) ("[T]he concrete nineteenth century man is for the most part an impure egoist, a mixed utilitarian.").

[31] Gary Becker, *Nobel Lecture: The Economic Way of Looking at Behavior*, 101 J. Political Econ. 385 (1993) ("[T]he economic approach I refer to does not assume that individuals are motivated solely by selfishness or material gain. It is a *method* of analysis, not an assumption about particular motivations. Along with others, I have tried to pry economists away from narrow assumptions about self-interest. Behavior is driven by a much richer set of values and preferences.").

[32] The analogy between theoretical models and maps plays an important illustrative role in the work of the eminent philosopher of science, Ronald Giere. *See generally,* Ronald Giere, Explaining Science: A Cognitive Approach (1988).

traditional economic method, it does beg the inquiry: how does the analysis change when the rational actor hypothesis is suitably nuanced to capture certain sorts of systematic departures from the rational actor model's predictions.[33]

Behavioral economics furnishes a panoply of enticing options, which might make a model of human behavior more "realistic."[34] However, for a number of complicated reasons, about I have written about elsewhere, I find pure behavioral theories objectionable. Thus, I shall take a somewhat more austere approach, more consistent with mainstream Law & Economics, preserving the basic framework of the Becker model, and supplementing it with a second order rational account of "bounded rationality."

3.2.1 Bounded Rationality

The behavioral economics movement was sparked by Herbert A. Simon's essay, *A Behavioral Model of Rational Choice*,[35] which introduced the concept of "bounded rationality" into the economic literature. The idea is that the very act of performing a cost-benefit deliberation is itself costly, and that deliberators are aware of this cost and work it into their deliberations—though standard cost-benefit analysis traditionally failed to account for it. Thus, in a world where information is not costless (as, indeed, it is not in the real world), the perfectly rational actor must devote all his time to accumulating background information and calculating what to do in order to make the "optimal" decision, which is almost certainly a *suboptimal* way of whiling away the day. Thus, Simon hypothesizes, what people *actually* do when deliberating is to set a threshold of acceptability, such that if some contemplated activity passes the threshold, it will be "good enough" to act upon, even if it is not the elusive "best" choice.

My model does not follow Simon as far as his hypothesis about thresholds of acceptability.[36] However, I do take decision costs and restricted decision-making domains as critical components of my model.

---

[33] It should be noted that while the rational actor model has been criticized for failing to predict enough human behavior to be a meaningful description, it remains largely undisputed that it is a compelling *normative* model of welfare-maximization.

[34] *See generally,* Christine Jolls, Cass R. Sunstein & Richard Thaler, *A Behavioral Approach to Law and Economics,* 50 Stanford L.R. 1471 (1998); Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 Science 1124-1131 (1974); Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decisions Under Risk,* 47 Econometrica 263-291 (1979).

[35] 69 Quarterly J. of Econ. 99 (1955).

[36] Simon refers to this phenomenon as "satisficing." Herbert A. Simon, *Rational Choice and the Structure of the Environment*, 63 Psychological Review 129 (1956) ("Evidently, organisms adapt well enough to 'satisfice'; they do not, in general, 'optimize.'").

The framework I propose is second order (although easily extendable to higher orders). Rather than considering the material costs and benefits of actions, my analysis considers the costs and benefits of how to decide on an action. The idea is that when actors regularly encounter similar factual circumstances, they have a variety of ways of deciding how to decide such problems. In some cases, where the stakes are very large, or where the fact pattern occurs infrequently, the actor may choose to perform a cost-benefit optimization. In other cases, where the decision costs are substantial relative to the difference in potential payoffs, actors will likely develop rules-of-thumb, cognitive "short cuts," as a way of automating decision-making, because the cost of case-by-case deliberation would represent a loss.

When selecting among the supernumerary possible cognitive shortcuts, it bears inquiring why an actor would choose one shortcut over another. The answer, which seems obvious on its face, is that he will select the shortcut that is welfare-maximizing, when decision costs are included in the calculation. However, incorporating decision costs is not as obvious as it may at first seem, since it begs the question why an actor would choose to second order *optimize*. The natural answer might be that second order optimization is third-order optimal. And so on, up the ladder.

There is some danger here that this "passing the buck" leads to an infinite regress. I will not attempt to offer a thorough answer to this deeply theoretical problem here; the infinite regress problem is tangential to the present topic.[37] Briefly however, it suffices to observe that we needn't consider higher and higher orders of heuristic-selection for the present inquiry, since the second order account seems a sufficiently plausible description to cover the interesting cases *sans* argumentative support.

Returning to the problem of criminal deterrence, there exists a salient characteristic, which we might well suppose rational actors will collectivize at a second order level: criminal activity. It may be that for many people, opportunities to commit criminal acts present themselves, which from time to time happen to be privately optimal. And yet, I suspect that in a majority of such cases, people decline to seize such opportunities, apparently choosing "suboptimal" law-abiding behavior and violating the rational actor hypothesis.[38] There exist a variety of philosophical explanations for such deviant virtuousness—personal morality being the most obvious one. On this view, actors internalize the norm that criminal act x is bad, and they get enough disutility from performing the act itself that it tips the cost-benefit balance toward

[37] *See* Daniel Pi, *A General Theory of Rationality* (2014) for a possible resolution of that problem.

[38] I admit that I may be naïve on this point, however it must necessarily be left to speculation, since it is difficult enough to estimate rates of undetected crimes, much less *unactualized* undetected crimes.

some other, presumably legally permissible activity.[39]

There may be something to the moral view. However, I contend that a second order rational explanation is more persuasive. If disincentives for crime are *generally* sufficient to deter the first order rational person from committing crimes, then (given that even petty crimes require some amount of planning and skillful execution) the second order rational person will simply adopt the second order heuristic not to commit crime. Thus, while such a second order rational person may from time to time miss out on golden opportunities to get away with a criminal act, he will refrain from thusly acting—not because of any compelling moral reason, but because such opportunities are so rare that they aren't worth contemplating. Consequently, they fall outside the domain of viable choices—outside the "bounds" of bounded rationality.

3.2.2 An Example

A simple example involving a petty crime will hopefully illustrate the point. Suppose that a first order rational person, call him Smith, parks his car on a particular metered stretch of road. He can either feed the meter or risk incurring a parking ticket. Let us assume that the penalty for parking illegally (i.e., without paying) is set at an efficient level à la Becker, such that the city has calculated the probability of catching a scofflaw, and set the fine at a rate, where repeated violations over time represent a losing gamble.

Smith, sufficiently motivated, could choose to stake out the road for several days, monitoring how frequently the meter-maids patrol the block, formulate a more fine-grained estimate of the probability of detection *on the particular block* at a given time, and decide whether to pay or not on the basis of that calculation. He could then "rationally" decide whether to pay or not, armed with a more refined estimate of probabilities, thus generating a net benefit over time.

Of course, we would not regard Smith's behavior as rational. Indeed, Smith is not only irrational, he is insane. Stalking meter-maids to accumulate sufficient information to generate an "optimal" decision comes at an *enormous* cost. Looking at the big picture, the information required to game the system surely represents a titanic net loss, and Smith's conduct would satisfy only the most myopic conception of "rationality."

Let us now consider a second order rational person, call her Jones. Jones has formed the heuristic that she will always pay for parking. From time to time, she will

[39] *See, e.g.,* Amartya Sen, *Rational Fools: A Critique of the Behavioral Foundations of Economic Theory*, 6 Philosophy and Public Affairs 317-344 (1977).

end up paying for parking, even when (unbeknownst to her) the expected benefit of not paying exceeds the expected cost. However, the opportunity cost being negligible, she merrily spends her deliberation efforts on more productive enterprises.

What the example makes clear, I hope, is that first order rationality often fails to be second order rational (it may also turn out, though less frequently, that second order rationality fails to be third-order rational, that third order rationality fails to be fourth-order rational, etc.). What is critical to observe is that the rational actor hypothesis (broadly construed) is not the source of absurdity, for the rational actor hypothesis is as true of Smith as it is for Jones.

Rather, to the extent that on some occasions Smith does not pay for parking, while Jones always does, the question is not whether Smith is irrational or Jones is irrational. Rather, they are *both* rational at different *orders* of decision-making. Thusly construed, the rational actor hypothesis—that actors make choices that maximize their private welfare—becomes a more general claim. And whereas first order rationality may seem absurd in some circumstances, higher order rationality seems likely to provide a more plausible account of our pre-theoretical understanding of the descriptor, "being rational." For convenience (and conformity to common usage), I will thus take "rational" to refer to the highest-order rationality within the scope of discussion.

3.2.3 Cashing Out the Public Policy

So how does this conception cash out as public policy? The proposition that if laws create sufficient disincentives, then rational actors will refrain from committing them, is hardly novel. Nor is the proposition that a subset of the population will be unresponsive to disincentives, for a variety of psychological and pecuniary reasons, which are difficult or impossible to address through the instruments of criminal punishment.

What is novel here is that enforcement need *not* be sufficient to effect disincentives for first order rational actors, since people do not typically behave first order rationally. Rather, the law should be designed to incentivize the formation of the second order heuristics, "not to park illegally," "not to steal cars," "not to burglarize," "not to download pirated software," etc. Indeed, an ideal criminal law regime would be generate the general heuristic, "not to commit crime." This entails a subtle, but significant shift in the way criminal law policy should be designed, since the target of incentives is only incidentally to create *direct* disincentives for criminal acts. Its primarily function, I contend, is to incentivize a *way of deliberating* about the commission of criminal acts.

The incentivization of "not to commit crime" heuristics may be accomplished in several ways. First and most obviously, enforcement should be targeted toward high-profile crimes. More visible enforcement is more likely to create the impression that the probability of detection is high, increasing actors' subjective assessments of detection probability, and thereby reducing their (subjective) expected payoffs. This is because the information cost for high-profile crimes is low (one need only read a newspaper to discover the latest criminal scandal of the day), whereas the cost of discovering the true rate of criminal detection and punishment is prohibitively high. Punishments for high-profile crimes should also be severe, for the same reasons.

Second, strategies should be developed to prevent the fragmentary formation of heuristics. For example, in the realm of automobile theft, the goal should be to encourage the development of the heuristic, "not to steal cars," and not the alternative heuristics, "not to steal expensive cars," or "not to steal new cars." If thefts of expensive or new automobiles are more aggressively investigated, then savvy criminals may be responsive to the poor payoffs involved with the theft of expensive or new cars, but form the alternative heuristic, "not to steal new or expensive cars," rather than the intended heuristic, "not to steal cars."

Asymmetric enforcement of crime may be inevitable, given limited police and prison resources. However, when the disparate enforcement of the law is obvious, actors may form alternative heuristics to exploit the asymmetries in enforcement, rather than adopting a wholesale "not to commit crime" heuristic.

In some sense, the second order analysis shifts the emphasis from enforcement-in-fact to enforcement-perception. Upon accepting such a conceptual shift, it becomes clear that the principal worry in criminal law is not that some people "get away" with crime, but rather that people will engage in particularized cost-benefit calculations whether to commit criminal acts. The more perceived exceptions and inequalities exist in the law, the greater the incentive to explore them in the hopes of "gaming" the numbers. The goal, I contend, of criminal punishment should not be to beat criminality at the first level, but to cut criminality off at the stage before it is even contemplated. The criminal law has not succeeded when a would-be criminal undertakes a cost-benefit analysis and is deterred by the expected cost of punishment—such an individual may still act criminally in some subsequent situation if the payoffs are right. Rather, the criminal law succeeds when a would-be criminal elects not to *consider* the option of committing crimes at all.

Upon reflection, this should not seem a daring proposition. A person who performs cost-benefit analyses whether to commit particular criminal acts, we may fairly suppose, does so because it is rational. The person who is always on the lookout for an opportunity to circumvent the law, if engaging in such criminal activity produces

a windfall, does so because the decision cost of *considering* whether to break the law is offset by the windfall benefits won by violating the law when he encounters a promising criminal opportunity. Common sense guides us well here: a person who is constantly peeking into parked cars, to see if someone left the door unlocked is a person who will at some point steal a car—even if he declines to steal *this* or *that* particular car. The most cost-effective strategy is not to lock as many cars as we can, but to stop the would-be thief from being on the alert for opportunities to steal cars.

It is my contention that if police and prison resources were better tailored to encouraging the formation of heuristics, rather than the brute prevention of crimes *via* legal disincentives, then the deterrent effect will be at least equal (and possibly better) than current criminal law practices. Moreover, the likely result would be a savings in police and prison costs and a mitigation of the trend toward overcriminalization. However, at this point, more precise machinery is required to establish those results concretely. And so I will now turn to the formal model in Section 3.3.

3.3 The Model

Let P_j denote the payoff function of a typical citizen, j. Let $B(x)$ represent the benefit of undertaking activity x, and let $[x \in] A_{j_n}$ represent the set of choices available to j in fact-situation n. Let $K(x)$ represent the cost of undertaking activity x. Finally, let $\pi(x)$ represent the probability of enforcement (i.e. the probability of detection, apprehension, and conviction, which we may further analyze as $\pi(x) = d(x) \times a(x) \times c(x)$); and $S(x)$ represent the magnitude of sanction associated with x.

Thus, the private payoff for parties contemplating criminal action will be the Beccarian function (Formula 3.1), and the first order optimal choice will be:

$$a^* = \max_x P_j(x) = B(x) - K(x) - \pi(x)S(x) \tag{3.4}$$

Now, let us consider second order welfare maximization. Let us represent decision-making process k with the function $D_k(F_n, A_{j_n}) = x$, which selects a possible action x from the set of possible actions A_{j_n} available to j, given a set of facts F_n. For example, welfare-maximization is one such decision-making process, call it D_R, such that $D_R(F_n, A_{j_n}) = \max_{x \in A_{j_n}} P_j(x)$. However, there may be an infinite number of decision-making processes, of which D_R is but one special example.

Let us now describe the second order payoff function. A heuristic is a decision-making process that is triggered when certain factual circumstances apply. Let us denote the triggering facts as $f_n \subset F_n$. Now, for any set of facts F_x, if it is the case

that $f_n \subset F_x$, then the heuristic will be applied. Thus, the second order payoff will be:

$$P_j^2(D_n, f_n, A_{j_z}) = \sum_{f_n \subset F_i} p(F_i) \left[(P_j(D_n(F_i, A_{j_z})) - \delta(D_n, F_i)) \right] \tag{3.5}$$

That is, the second order payoff of decision procedure D_n is the sum of first order expected payoffs when using decision procedure D_n, minus the decision cost of D_n, which we denote $\delta(D_n, f_y)$, discounted by the probability that the situation F_i occurs. Thus, the second order optimal choice of decision-making procedure for a given fact-trigger f_y will be:

$$D^* = \max_{D_x} P_j^2(D_x, f_y, A_{j_z}) \tag{3.6}$$

Before proceeding, it is worth observing that the optimal decision-making procedure may fail to be unique, depending on the subset of triggering facts f_y. Two heuristics may develop, which are both second order optimal, but which overlap. That is, one heuristic may be triggered by facts f_y, while another heuristic may be triggered by facts f_z. Particular situations may arise, where some total set of facts F_w describing an actual circumstance is such that both triggers are activated, $f_y \subset F_w \wedge f_z \subset F_w$.

For example, suppose someone were to form the heuristic, "Do not trust X," with the triggering fact "X is a Cretan'" And further suppose that the same person forms the heuristic, "Trust Y," with the triggering fact "Y is a philosopher.'" Certainly, there will arise a conflict of heuristics, when the person encounters a philosopher from Crete.

However, this is not a problem for the model. When a situation triggers multiple heuristics, which generate contradictory choices, other heuristics will be implicated to resolve the conflict of heuristics. In particular, for two i^{th}-order heuristics with overlapping triggering conditions, the $(i+1)^{\text{th}}$-order decision-making procedure that led the actor to adopt them should provide a basis for deciding which one trumps and when.[40]

Much more could be said about resolving the problem of conditions that trigger multiple heuristics, however this would take us far afield of the present inquiry (and indeed would require a generalized account of higher order rationality), but it suffices merely to observe that this is not problematic for the theory in our application here.

[40] If the two heuristics were adopted due to different $(i+1)^{\text{th}}$-order decision-making procedures, then the $(i+2)^{\text{th}}$-order decision-making procedure that led the actor to adopt the two $(i+1)^{\text{th}}$-order decision-making procedures should decide which trumps and when, and so on.

Proposition 3.3.1. *In the absence of decision costs (i.e., $\delta = 0$), $D^* = D_R$.*

Proof. See Appendix. □

Proposition 3.3.1 states the intuitively obvious baseline, where the second order optimal decision-making procedure will be first order rational deliberation when decision costs are ignored. This proposition explains why the first order rational actor hypothesis is thought to represent a normative ideal.

Proposition 3.3.2. *The product of a second order optimal decision-procedure will not always be coextensive with first order optimal decisions, $\neg\forall a^*(a^* = D^*(F_n, A_{j_n}))$.*

Proof. See Appendix. □

Proposition 3.3.2 is a critical point. Therefore, it may be worthwhile to furnish an example to illustrate. Consider the potential car-thief, who peeks into the windows of parked cars, checking to see whether they've been left unlocked, whether they have security alarms, whether they have LoJack, whether they have steering wheel locks, etc. Say that he performs a cost-benefit calculation for each car. Let us assign some hypothetical values, to see how a decision that is first order optimal may fail to be second order optimal.

	Benefit-Cost	Enforcement Prob.×Sanction	Deliberation Cost
Car 1	100-80=20	$.05 \times 1000 = 50$	1
Car 2	109-70=39	$.04 \times 1000 = 40$	1
Car 3	105-90=15	$.05 \times 1000 = 50$	1
Car 4	120-85=35	$.06 \times 1000 = 60$	1
...	...	...	...
Car 49	90-70=20	$.04 \times 1000 = 40$	1
Car 50	100-25=75	$.06 \times 1000 = 60$	1

Let the probability of enforcement vary, depending on whether the car is parked in a private lot, or whether it is parked on the street, whether the surrounding area is high-traffic, well lit, etc. Suppose the punishment for automobile theft creates a disutility value of 1000. And suppose that for all but one of the cars, there exist various security features, which render theft of the vehicle a net loss for the thief. But notice that Car 50 represents a potential gain—perhaps because the owner left the car unlocked, rendering the cost of undertaking the theft relatively low. Thus, the payoff of stealing Car 50 is positive.

Now, on the first order Becker-style analysis, the law has—in the particular case of Car 50—failed to set sufficiently high disincentives (note, this may still be an "efficient" level on Becker's account). The value of the car, less the effort to steal it, less the expected cost of enforcement, and less decision cost is $75 - 60 - 1 = 14$. Assuming opportunity cost $OC < 14$, the thief will choose to steal Car 50.

But does the thief really come out ahead? Of course he does not, because even though he made the rational decision *not* to steal Cars 1-49, each rational calculation cost him 1 in deliberation. Thus, after the whole exercise, the (first order rational) thief comes out behind, $14 - 49 = -35$.

Contrast the rational thief with a person, who has formed the heuristic "not to steal cars." Her per-decision decision cost will be zero, and her choice in every case will simply be not to steal the car. It is true that she "misses out" on the potential surplus of stealing Car 50. However, she comes out ahead overall, since her total welfare from adopting a "not to steal cars" heuristic will be 0, as compared with the first order rational thief, whose welfare is -35. Thus, the optimal decision-making procedure can, given certain values, generate sets of decisions, which are in the aggregate more optimal than the "optimal." A startling result.

Let us now consider what happens when we reduce the sanction from 1000 to 900. Becker predicts that decreasing the "price" of crime will increase consumption, leading to more cars being stolen.

	Benefit-Cost	Enforcement Prob.×Sanction	Deliberation Cost
Car 1	100-80=20	$.05 \times 900 = 45$	1
Car 2	109-70=39	$.04 \times 900 = 36$	1
Car 3	105-90=15	$.05 \times 900 = 45$	1
Car 4	120-85=35	$.06 \times 900 = 54$	1
...	...	...	...
Car 49	90-70=20	$.04 \times 900 = 36$	1
Car 50	100-25=75	$.06 \times 900 = 54$	1

According to Becker's theory, by reducing the severity of punishment from 1000 to 900, the "price" of crime is decreased, and the result will be an increase in the activity level of criminals. In particular, it seems that Car 2 now represents a net profit of 2, so another car will ostensibly be stolen, due to the reduction in sanction, assuming zero opportunity costs.

On my view, the reduction will *not* necessarily result in an increase in car thefts, because it remains second order rational to adopt the heuristic "not to steal cars." That is, even with the improvement in the expected payoff $2 + 21 = 23$, the decision

costs (-50) continue to represent a net loss (-27), while the heuristic "not to steal cars," remains at 0. Thus, the second order rational person will continue to prefer not to bother contemplating car theft, even though the potential for a windfall gain has slightly improved.

Thus, despite a 10% reduction in punishment, the effect of deterrence may well remain constant. This result does not entirely contradict Becker's theory, for price theory effects do play a role, even in my second order rationality conception. For example, for individuals, whose opportunity cost is $-35 < OC < -27$, it will be privately optimal to rationally weigh the costs and benefits of attempting a particular theft, and in the examples above, Car 2 will be stolen if sanctions are decreased to 900. However, it is curious how a person could have such low opportunity costs (indeed, it seems that *doing nothing* ordinarily has a payoff of 0, so that ordinarily $OC \geq 0$).[41] And it remains true that a sufficient diminution in the price of crime will ultimately increase consumption (in the extremum case, observe that when $S(x) = 0$, the heuristic people will form will be to *always* steal cars when possible). However, the foregoing example illustrates that crime rates may be considerably less elastic (with respect to enforcement) than under the Becker model, and that up to a point, enforcement may be reduced without any reduction in deterrence effects.

Proposition 3.3.3. *Assuming actors are second order rational, maximal deterrence will be achieved at the point where it becomes second order rational to adopt the heuristic "not to commit crime."*

Proof. See Appendix. □

The idea here is that assuming actors are rational, the maximal level of deterrence will be achieved when parties choose to adopt a heuristic not to commit crime. The alternatives would be the adoption of, for example, the heuristic, "not to steal locked cars," or "not to steal cheap cars," or "not to steal expensive cars," or "not to steal red cars." Such alternative heuristics may reduce decision costs, so that the net private gains represent an improvement over the "not to steal cars" heuristic.

Practically, one mechanism for combatting the formation of opportunistic heuristics, which exploit asymmetries in enforcement, would be to create increased second order decision costs. For example, if police resources are limited, making it unfeasible to pursue all car theft cases effectively, resources will have to be focused. Some

[41] It is possible that $OC < 0$ if for example, a prisoner of war is being tortured, and must weigh whether to attempt an escape. In that case, the cost of doing nothing may be negative, such that even a long-shot attempt will still be optimal. One must admit, however, that such scenarios are rare.

cases will go by the wayside, so greater resources can be spent on effectively pursuing a targeted subset of cases. In deciding which cases to investigate, and which cases to "ignore," it would be a mistake to focus resources on high value targets or low value targets. Instead, resources should be allocated in a way that would be difficult for would-be criminals to discern. For instance, investigating car thefts, where the stolen vehicle's license plate ends in an even number from January through June, and investigating stolen vehicles, where the license plate ends in an odd number from July through December. Certainly, if car thieves knew about the policy, they could easily exploit it. However, the decision cost involved in generating such a heuristic would be enormous, since in the absence of an informant with "inside information" about police practices, car thieves would have to suss out the necessary information through trial and error and careful analysis of the data.

Certainly, asymmetric enforcement is an inevitable consequence of limited resources, however when asymmetric enforcement falls along obvious lines, it "helps" criminals by reducing second order decision costs and serving up easy-to-follow heuristics to avoid detection. When asymmetries in the allocation of enforcement are unavoidable, the distribution of enforcement resources should be calculated to employ decision costs as a tool to "hide" the asymmetries, such that only an industrious econometrist would be capable of discerning the circumstances where the commission of crime entailed an expected gain.

Proposition 3.3.4. *Inducing the development of a "not to commit crime" heuristic is less costly than optimal first order deterrence.*

Proof. See Appendix. □

Proposition 3.3.4 expresses the comparative static that incentivizing the development of a "not to commit crime" heuristic is more cost-effective than Becker's formulation of optimal deterrence. The practical cash-out is that if deterrence is the goal of criminal law, then we may be spending more than necessary—both on enforcement and punishment—and that *way* we invest those resources is also inefficient.

3.4 Practical Considerations

I am hesitant to speculate about how the framework I have presented will pay out in concrete policy terms. Such conclusions would require both a theory and real-world data, of which I have only offered the former. It is beyond the scope of this paper (and my competence) to offer the latter. However, it may be worth suggesting some

common-sense hypothetical applications of my theory, which at least provisionally point the way toward possible real-world implementation.

First, if the reason why most ordinary people do not ever commit felonies is because it is second order optimal for them to form a "not to commit crime" heuristic, then a substantial majority will continue to refrain from serious criminal activities even if the severity of sanctions and rigorousness of policing were decreased. One consequence of my theory is that the "consumption" of crime is substantially less elastic than under Becker's model, and therefore crime rates will remain relatively stable, with respect to changes in enforcement levels. The optimal points of detection effort and sanction predicted by my model will therefore be lower—possibly *much* lower—than that predicted by Becker.

Determining precisely *how much* less enforcement will effect "optimal" deterrence is a difficult empirical question. However, if I am also right that the deterrence effect given present enforcement levels in industrialized nations has passed the point of diminishing returns, then it may well be that for the $< 10\%$ of the population who commit at least one felony offense in their lives, inducing the formation of a second order heuristic is either impossible or prohibitively costly. Thus, for the $> 90\%$ of law-abiding citizens, policing and sanctions may be far in excess of what is required to induce the second order heuristic "not to commit crime," and significant reductions in enforcement may have no detectable effect on crime rates.

More practical still, in lieu of empirical research on this point, policymakers can discover what the threshold and optimal points of deterrence are by gradually reducing sanction severity up to the point where measurable increases in criminal activity are detected. That is, we do not necessarily need studies to get to the desired policy—this can be done through simple trial and error. Reducing punishment levels and policing will also have the ancillary benefit of reducing the supernumerary social injustices associated with the penal system,[42] and result in pecuniary savings for the state.

Moreover, not only can enforcement costs be reduced, but enforcement can also be made more effective. If the narrative that my model tells about law-abiding citizens is that they form a second order optimal first order heuristic "not to commit crime," and criminals tend to be resistant to the formation of such heuristics, then we can use this information to expand the reach of deterrence to the $< 10\%$ who have hitherto been unresponsive to deterrence incentives.

For example, we may be better able to instill general "not to commit crime" heuristics by intervening earlier on in the process of heuristic formation: focusing resources on children and adolescents, increasing policing and sanctions for petty

[42] *See generally* Douglas Husak, Overcriminalization (Oxford University Press 2008).

crimes, and reducing poverty (thereby increasing the opportunity cost of criminal activity—though this strategy would also be effective under Becker's conception, the responsiveness of individuals to increasing opportunity costs is more elastic in my conception, thus the marginal benefit of poverty-reduction would be somewhat greater. Of course, these suggestions are hardly novel. However, my model and alternative policy objective suggest that they may be more effective than traditional economic models would predict, suggesting a different allocation of resources. Moreover, working from a heuristic-formation perspective will sharpen the policy goals of juvenile justice reform, "broken windows" policing,[43] and social welfare programs. It also supplies an economic argument for further investment in such policy goals.

There is also room for much creativity in designing policies around information costs. For example, several of former New York City mayor Rudolph Giuliani's actions during his time as the U.S. Attorney for the Southern District of New York seem likely to have economized on decision costs (albeit most likely inadvertently). For instance, Giuliani famously favored public arrests of high-profile individuals, which attracted much media attention, even though the charges were often later dropped or reduced.[44] To be sure, this strategy has obvious defects: arguably violating the rights of the dubiously humiliated high-profile figures, and risking decreased public confidence in the rule of law. However, the idea of "selling" the notion that white-collar criminals (or earlier in Giuliani's career, mafia bosses) are as vulnerable to enforcement as other citizens surely exploited cheap publicity for the message that the expected payoff for criminal activity is negative. Giuliani also reputedly instituted a policy of aggressively prosecuting different types of crimes exclusively on certain days of the week and neglecting those that were not the arbitrarily chosen prosecution-du-jour.[45] This would effectively "hide" the asymmetries in enforcement by increasing the decision cost of discovering what crimes would actually be (aggressively) enforced, and thereby frustrate criminals looking to exploit asymmetries in legal enforcement (the asymmetries of course were still present—limited resources necessitates some sort of distribution of enforcement efforts—the point is that assigning particular offenses to arbitrary days of the week had the effect of making those asymmetries more difficult to perceive and thus more difficult to exploit, discouraging fragmentary heuristic-formation).

Finally, it bears highlighting a critical point, which may be easily confused. I am

[43] George L. Kelling & James Q. Wilson, *Broken Windows: The Police and Neighborhood Safety*, Atlantic Magazine (March 1982).

[44] Joel Cohen, *National Law Journal* (August 5, 2002).

[45] This story may be apocryphal, but it does not much matter whether Giuliani actually adopted such a policy—it suffices to point out that such a policy would be favored under my model.

not arguing simply for an increase in information costs. If the cost of information were increased, then under a trivially modified version of Becker's theory, criminals would decline to attempt crimes simply because they have become more costly. Rather, my point is that high information costs trigger heuristics, and that by distributing information costs in a certain way (uniformly if possible or unpredictably if resources are limited), would-be criminals are not deterred because the increased cost of information has made the commission of a particular crime unprofitable, but rather because the aggregate cost of *rational deliberation* has become unprofitable. Even if Becker's model were tweaked to account for information costs, it would still yield different predictions from the model I have proposed here.

3.5 Conclusion

There are several insights I have hoped to communicate in this essay. First, I hope to have contributed a methodological novelty, suggesting an alternative role for behavioral economics in social engineering—the opposite of debiasing: "biasing." As opposed to situations where first order rational behavior is the normative goal, and where irrational biases and heuristics present obstacles, in the realm of criminal law, first order rational behavior may sometimes be the source of the problem. In such cases, the law should encourage citizens to develop "irrational" biases, which lead them away from self-interested rational behavior.[46] The use of biases and heuristics therefore presents us with a new tool in the policymaker's toolkit.

Second, I hope to help bridge the divide between the economic theory of criminal law and criminal law sociologists. I concede that further research will be required to determine whether my theory is predictive in fact, however it is at least a plausible framework for making the economic account of criminal deterrence more compatible with the empirical data gathered by social scientists.

Third, I have identified a more nuanced policy goal for the criminal law: encouraging the formation of a "not to commit crime" heuristic rather than merely effecting first order deterrence for particular cases. And I have suggested several ways of pursuing this new objective. Developing further strategies for generating anti-crime heuristics promises to be a fecund new territory for future research.

[46]This is different from traditional incentive-alignment, which has typically assumed that actors are first order rational. My methodological contribution will have been to show how inducing systematic non-rational behavior can serve as an additional mechanism for incentive-alignment.

3.6 Appendix

Proposition 1. *In the absence of decision costs (i.e., $\delta = 0$), $D^* = D_R$.*

Proof. By definition, $P_j(D_R(F_n, A_{j_n})) = P_j(a^*) = P_j^*$. Thus, from Formula 3.5, we know that $P_j^2(D_R, f_n, A_{j_n}) = \sum_{f_n \subset F_i} \left[p(F_i)P_j^*\right]$, assuming $\delta(D_R, F_x) = 0$. From Formula 3.6, we know that $D^* = \max_{D_x} P_j^2(D_x, f_y, A_{j_z})$, and because of Lemma 3.6.0.1, it follows trivially that $D^* = D_R$. □

Lemma 3.6.0.1. $\max_{X=\langle x_0,x_1,...,x_n\rangle} \sum_{i=0}^{n} g_i(x_i) = Z$, *such that* $Z = \langle z_0, z_1, ..., z_n\rangle = \langle \max_{x_0} g_0(x_0), \max_{x_1} g_1(x_1), ..., \max_{x_n} g_n(x_n)\rangle$.

Proof. Suppose for *reductio* that $\max_{X=\langle x_0,x_1,...,x_n\rangle} \sum_{i=0}^{n} g_i(x_i) \neq Z$. This implies that there exists an ordered set $Z^\dagger$, such that $\max_{X=\langle x_0,x_1,...,x_n\rangle} \sum_{i}^{n} g_i(x_i) = Z^\dagger = \langle a_0, a_1, ...a_n\rangle$ and that $\exists a_y \exists z_y (a_y \in Z^\dagger \neq z_y \in Z)$.

We now proceed by mathematical induction (assuming that g_i is independent of g_{i+1}). If $a_0 \neq z_0$, then $\sum_{i=0}^{0} g_i(z_i) > \sum_{i=0}^{0} g_i(a_i)$, since $z_0 = max_x g_0(x)$. Therefore, it must be the case that $a_0 = z_0$. And likewise, if $\sum_{i=0}^{k} g_i(z_i) = \sum_{i=0}^{k} g_i(a_i)$ and $a_{k+1} \neq z_{k+1}$, then $\sum_{i=0}^{k+1} g_i(z_i) > \sum_{i=0}^{k+1} g_i(a_i)$, since $z_{k+1} = max_x g_{k+1}(x)$. Therefore, $\sum_{i=0}^{k+1} g_i(z_i) = \sum_{i=0}^{k+1} g_i(a_i)$. By induction, this proves that $Z^\dagger = Z$, contradicting the assumptions that $\max_{X=\langle x_0,x_1,...,x_n\rangle} \sum_{i=0}^{n} g_i(x_i) \neq Z$ and that $\max_{X=\langle x_0,x_1,...,x_n\rangle} \sum_{i}^{n} g_i(x_i) = Z^\dagger$. □

Proposition 2. *The product of a second order optimal decision-procedure will not always be coextensive with first order optimal decisions, $\neg\forall a^*(a^* = D^*(F_n, A_{j_n}))$.*

Proof. Suppose that $\sum_{f_n \subset F_i} p(F_i)(P_j(D_R(F_i, A_{j_z})) = r$, and that $\delta(D_R, f_n) = e$. And suppose there exists some alternate decision-making procedure D_Q, such that $\sum_{f_n \subset F_i} p(F_i)(P_j(D_Q(F_i, A_{j_z})) = q$, and that $\delta(D_Q, f_n) = h$. It follows trivially that if $r - q < \sum h - e$, then in some cases, the non-first order rational choice $\exists x \exists y (y = D_Q^*(F_n, A_{j_z}) \wedge x^* = D_R(F_n, A_{j_z}) \wedge x^* \neq y)$. □

Proposition 3. *Assuming actors are second order rational, maximal deterrence will be achieved at the point where it becomes second order rational to adopt the heuristic "not to commit crime."*

Proof. This proof follows trivially from the definitions. In the interest of thoroughness, however: Let D_{NC} denote the heuristic "not to commit crime." Clearly then, $\delta(D_{NC}) = 0$, regardless of the inputs; and for any $x = D_{NC}(f_n, A_{j_n})$, where f_n contains the fact that the contemplated act is criminal (and only that fact), the output

x will be to decline to commit the crime in question, which presumably generates the private benefit 0.[47]

Let f_a include the fact that the contemplated act is criminal *and* some other factor ω, and let f_b include the fact that the contemplated act is criminal *and* the factor that $\neg\omega$. Thus, if $S = \{F_i : f_a \subset F_i \vee f_b \subset F_i\}$ and $T = \{F_i : f_n \subset F_i\}$, then $S = T$.

Let D_{A_1} and D_{A_2} be alternative heuristics, such that $D_{A_1}(f_a, A_{j_a})$ is to decline to commit the crime, with decision cost $\delta(D_{A_1}) = 0$; and $D_{A_2}(f_b, A_{j_b})$ may (or may not) prescribe undertaking the criminal act, with decision cost $\delta(D_{A_2}) = \mu$.

Clearly, if D_{A_2} ever prescribes undertaking a criminal act, the combination of D_{A_1} and D_{A_2} will be less than maximally deterrent, because D_{NC} would decline to undertake that criminal act. If D_{A_2} never prescribes undertaking a criminal act, then the combination of D_{A_1} and D_{A_2} are maximal, but extensionally equivalent to D_{NC}, though possibly with higher decision costs, where $\mu > 0$. Thus D_{NC} is the maximally deterrent heuristic, though possibly not uniquely. □

Proposition 4. *Inducing the development of a "not to commit crime" heuristic is less costly than optimal first order deterrence.*

Proof. If we assume first order rationality, then the necessary enforcement required to deter potential criminals from committing a potential crime (in a particular situation F, such that the factual circumstances f exist to commit a crime, i.e., $f \subset F$) will be:

$$\pi^1 S^1 = B - K - \Gamma + \epsilon$$

where Γ is the opportunity cost, and ϵ is some "kicker" to effect $\pi^1 S^1 > B - K - \Gamma$.[48] For second order rationality, the necessary enforcement required to deter potential criminals from committing a potential crime will be:

$$\sum_{f_n \subset F_i} \left(\pi^2 S^2\right) \geq \sum_{f_n \subset F_i} (B - K - \Gamma - \epsilon - \delta)$$

It is easy to see that, *ceteris paribus*, $\pi^1 S^1 > \pi^2 S^2$: first consider if the *average* payoff for a type of crime h were $B_h - K_h - \Gamma_h$, and if the average cost of rational deliberation for such a crime $\delta_h(D_R, f_n \subset F_i) > 0$. Trivially then, it would be the case that:

$$\pi_h^1 S_h^1 = B_h - K_h - \Gamma_h - \epsilon > B_h - K_h - \Gamma_h - \delta_h - \epsilon = \pi_h^2 S_h^2$$

[47] It may be worth observing here that such "negative heuristics" may be better understood as constraining A_{j_n} rather than prescribing the "null" action.

[48] I will use superscripts $\pi^n S^n$ to denote sufficient enforcement at n-order to deter crime.

Moreover, if $\pi_h^2 S_h^2 > B_h - K_h - \Gamma_h - \delta_h$ for a set of triggering facts f_n, then according to my model, potential criminals will adopt a "not to commit crime (with triggering facts f_n)" heuristic, and thus $\pi_h^2 S_h^2 = \pi^2 S^2$, whereas under the first order rational model, there may be outlier cases k, such that $\pi_k^1 S_k^1 > \pi_h^1 S_h^1$. To achieve the same result as a "not to commit crime" heuristic, first order deterrence theory would suggest the necessary deterrence level should be set at $\pi^1 S^1 = B_j - K_j - \Gamma_j$, where $B_j - K_j - \Gamma_j$ represents the maximum possible surplus derived from committing a type of criminal act.

Thus, the required enforcement under a first order deterrence model will be far higher than under a second order deterrence model. And trivially therefore $C(\pi^1, S^1) > C(\pi^2, S^2)$. □

Chapter 4

Harmful Speech

The freedom of expression is a defining characteristic of liberal democracies.[1] In practice, exceptions are invariably allowed to pervade, which permit putatively open societies to regulate disfavored categories of speech. A predominant justification given for such regulation is that some classes of expressive activity are supposed on balance to be *harmful*. This article presents an argument undermining that justification. I contend that merely identifying a category of speech as generating low social benefit and high social cost is an insufficient ground to justify regulation of that speech.

The arguments I oppose are methodologically consequentialist and economic in nature. I correspondingly formulate my counterarguments to the received view within that same analytical framework. Certainly not all arguments for speech regulations are of this kind. Alternatives exist. For example: that the law should be an instru-

[1]Indeed, even the most brutally *illiberal* governments have at least *nominally* recognized a general speech right. For example, presumptions against the regulation of speech were enshrined in the constitutions of Nazi Germany, *Weimar Constitution* art. 118 (1919), the Soviet Union, *Constitution of the Russian Socialist Federated Soviet Republic*, art. 2.14 (1918), Maoist China, *Constitution of the People's Republic of China*, art. 87 (1954), and North Korea, *Constitution of North Korea*, art. 67 (1972, rev. 1998). Although Libya under Gaddafi lacked a formal constitution, it was nevertheless the official position of the Libyan government that citizens should enjoy unlimited expressive rights. Gaddafi, Muammar. *The Green Book* (1975) ("An individual has the right to express himself or herself even if he or she behaves irrationally to demonstrate his or her insanity."). Indeed, Libyan law extended full freedom of expression even to commercial speech, even specifying tobacco advertising as an example where speech rights trump public health interests. However, Gaddafi also proclaims, "private individuals should not be permitted to own any public means of publication or information," *id.*, which effectively rendered the right irrelevant. Moreover, even residual "non-public speech rights" tended to be disregarded in practice in Gaddafi's Libya—as in the aforementioned authoritarian regimes.

ment for preserving a community's "cultural identity," or that a class of expressive activity is "contrary to the will of to God." This article are not aimed at addressing those alternative rationales. I take harm-based justifications to be the best and most frequently advanced bases for speech regulation, and rebuttal of that class of justification is the sole object of the present inquiry.

My exposition consists in three main parts. In the first, I provide a generalized construction of the harm-based justification for speech regulation and establish its prevalence in jurisprudential and scholarly thought. Next, I identify three countervailing effects, which undermine the logic of the received view. In the presence of these countervailing effects, I demonstrate that government efforts to reduce harmful speech will be less effective than proponents of speech regulation have supposed, and that such regulations may even exacerbate the very harms they were intended to remedy. Finally, I sketch out how the three effects may arise in specific speech contexts.

4.1 The Argument *for* Harm-Based Regulation

Let us begin by abstracting the class of speech restrictions which limit an individual's right to expression in cases where it generates negligible social value and imposes a substantial negative externality. This is definitional. I take these two conditions to *define* "harmful speech," demarcating the scope of the present inquiry.

Definition. *"x is harmful speech" is true if and only if:*

(condition i). *x generates negligible social value, and*

(condition ii). *x imposes a substantial negative externality.*

The two conditions are stronger than necessary for my argument to work. The principles I exposit below are generally applicable *whenever* social welfare maximization is used to justify speech regulation. In other words, the arguments I advance in Section 4.2 apply even if "harmful speech" were simply defined as "net social welfare reducing speech." Nevertheless, for practical purposes it is useful to constrain the definition of "harmful speech" in terms of conditions (i) and (ii), because these are the terms in which courts and scholars have tended to formulate their justifications for speech regulation.

Presumably, the reasons why courts and scholars have limited speech regulations to instances where conditions (i) and (ii) are satisfied (rather than *any* net welfare-reducing speech) are preemptive. They anticipate several common objections to

speech regulation, with which I agree, but which I do not investigate in depth here. First, that it is exceedingly difficult to quantify and compare the value and social cost of speech. Second, that the risk of "slippery slope" effects abound. Third, that citizens of liberal democracies derive utility from the very possession of the legal right itself—beyond its merely instrumental value. In light of these reasons, harm-based rationales for speech regulation typically assume that the social costs of a class of expression should not merely be greater than its benefits, but *substantially* greater than its benefits before the government is justified in interfering. For this reason, I formulate "harmful speech" in terms of "negligible" social value and "substantial" negative externality. Of course, this does not lessen the burden for my position. To the contrary, it distills the claims of the received view to its strongest cases. The principles exposited below follow *a fortiori* for all harm-based justifications for speech regulation.

Next, for present purposes I define "speech regulation" as being either the punishment or prior restraint of expressive activities. The term "regulation" is, of course, sometimes understood more broadly than this.[2] I consider the subsidization of speech only in passing. When I use the term "speech regulation," I exclusively mean either ex ante restraint or ex post punishment of expressive activity. Within the ambit of ex post punishment, I include all sanctions which would decrease the utility of the speaker engaging in the proscribed speech. This includes both criminal and civil liability.

Next, observe that arguments for the limitation of speech rights *assume* the existence of a general speech right. It is not sensible to speak of justifying a limitation of legal rights which are not recognized as rights in the first place. In the United States, the general speech right is established in the First Amendment of the Constitution.[3] The extent of the general right is unspecified in the text,[4] and it is generally understood to be the prerogative of the judicial system to determine its boundaries.

Broadly, courts have adopted two approaches to limiting the speech right. First, they can deem a category of behavior to fall outside the ambit of "speech." In other words, courts can decline to legally recognize an activity as *being speech*, though it may have some incidental expressive component. Therefore, being non-speech, it falls outside the protection of the First Amendment. The government can thus

[2]For example, "regulatory activities" may broadly be taken to include subsidies also.

[3]U.S. Const. amend. I.

[4]There is a reasonable textualist argument that the extent *is* specified. "Congress shall make no law . . . abridging the freedom of speech, or of the press," implies that *any* law abridging speech violates the rights of citizens. *Id.* Courts have declined to accept this view, and the "no law" language is customarily treated as something less than absolute.

regulate it—assuming of course that the proposed regulation otherwise falls within the powers which the Constitution assigns to the government. Second, the courts can acknowledge a category of behavior as *being speech* properly, but hold that public policy demands an exception.[5] The most forceful public policy rationales typically involve the identification of some *social harm*, which the regulation of speech is meant to remedy.

There is little difference between the two approaches in effect. Whether the courts declare a category of conduct to be "non-speech," or "speech for which an exception to the general right exists" is a merely formal distinction. It is immaterial in practical effect. The justification for adopting either approach—to the extent that the justification is grounded in the harmfulness of a speech category—will employ essentially identical reasoning.

Suppose the government proposes to regulate a class of expressive activity. Call it x. Harm-based arguments advocating for the regulation of x share the general form:

---

[5]The battle lines are somewhat less clear than I have indicated when viewed up close. For example, the "traditionally exclusions" of the general speech right are sometimes argued to be non-speech, and sometimes argued to be speech but within a recognized exception. These "traditional" categories include defamation, obscenity, and incitement. The distinction is important in the structure of the argument. If the categories are defined as falling *outside* speech, then no justification for their regulation is needed. However, if they fall within the ambit of "speech," then their regulation *does* require justification. To the extent the latter view is held, the arguments I present in this article will apply.

The arguments given for regarding such expressions as non-speech are often "originalist" in nature. The originalist argument is that these suspect speech categories were assumed to be permissible in the English law, and that the framers of the Constitution would not therefore have regarded nor intended the First Amendment to be a bar on their regulation.

This is clearly mistaken. James Madison, addressing Congress on the adoption of the Bill of Rights, expressly sought a break from past practice, distinguishing the guarantees of individual liberty from earlier political and legal norms, saying, "But although . . . it may not be thought necessary to provide limits for the legislative power in [Britain], yet a different opinion prevails in the United States," and "The right of freedom of speech is secured; the liberty of the press is expressly declared to be beyond the reach of this Government . . . ." 1 Annals of Cong. 436, 738 (1789).

Madison does *not* say, "*unless* of course the Government's interest were *compelling*, in which case, infringe away!"

Yet lest I be seen to endorse amateurish historical speculation, I should add that I am skeptical what value there is in the originalist mode of Constitutional interpretation, that *I* do not pretend to be a historian, and that I question how seriously we should take the moral speculations of bewigged, pre-Industrial-Revolution plantation owners. I point to Madison's remarks merely to indicate an "even if" argument. Regretfully, a fuller discussion of the topic would require a digression of such magnitude as to overwhelm my present thesis.

1. x is harmful speech.

2. If harmful speech is punished, then social welfare is improved.

3. If x is punished, then social welfare is improved. (*from* 1, 2)

4. If punishing x improves social welfare, then the government should be permitted to punish x.

5. Therefore, the government should be permitted to punish x. (*from* 3, 4)

The target of my counterargument is premise (2). If it can be shown that (2) is false, then the inference to (3) fails. And without (3), the inference to (5) fails. And if the inference to (5) fails, then merely establishing the fact that some x is harmful speech will be insufficient to justify regulation of x. This is the critical point: interrupting the inference to (5).

It bears remarking that premise (4), as I have formulated it, is also vulnerable to counterargument. Concededly, it could be nuanced to better reflect the varying levels of judicial scrutiny which the courts have applied to First Amendment controversies. Regardless, premise (4) is not the target of my present inquiry, and if I have stated (4) too extravagantly, this does no damage to my counterargument, for I stipulate the point *arguendo*. Again, stipulating to (4) does not lessen the burden of my argument, but rather strengthens the argument I oppose.

Of course, proponents of speech regulations do not assume premise (2) axiomatically. It is supportable with further argumentation. For convenience, let us call the supporting sub-argument the "harm-deterrence argument":

2.1. If harmful speech is punished, then the supply of harmful speech decreases.

2.2. If the supply of harmful speech decreases, then social welfare is improved.

2. If harmful speech is punished, then social welfare is improved. (*from* 2.1, 2.2)

This can of course be analyzed still further. I expect premise (2.1) to be founded on a model of deterrence, borrowed from the standard economic analysis of criminal law.[6] And premise (2.2) seems to follow trivially from the definition of "harmful speech" (conditions (i) and (ii) above).

[6] *See* Gary Becker, *Crime and Punishment*, 76 J. Pol. Econ. 169 (1968).

The harm-deterrence argument has a kind of elegance, benefitting from its resemblance to the economic approach to criminal and tort law.[7] By association with that scholarship, it *feels* familiar, uncomplicated, and vaguely "right."

Of course, the explicitly economic reconstruction of the argument for harm-based regulation is seldom stated in the general form I have articulated. The express formualtion is however not entirely absent in the scholarly literature. Judge Richard Posner, ever the pioneer, attempted a general economic analysis of speech rights in *Free Speech in an Economic Perspective*. Posner devises a cost-benefit framework for interpreting the First Amendment, which is essentially equivalent to the harm-deterrence argument above.[8] Unfortunately, Posner's initial push, which he carefully distinguished as being "partial" and "tentative,"[9] has not succeeded in inspiring many economists to expand upon his work. And in the three decades since its publication, Posner's article remains still the most comprehensive treatment of speech rights from an overtly economic perspective.

It would however be a mistake to characterize the harm-deterrence argument as *Posner's* argument. The policy rationales of *non*-economists used to justify speech regulation are very nearly all *implicitly* economic justifications, which tacitly rely upon the harm-deterrence argument for their foundation. Indeed, *most* of the frequently rehearsed arguments for speech regulation are really harm-based economic justifications in disguise. This last assertion, of course, requires some defense.

To foreclose potential complaints that I am setting up a straw man, some examples of courts and legal scholars utilizing the harm-based justification for speech regulation are wanted. The remainder of this section addresses this concern.

I aim to show in subsections 4.1.1–4.1.7 that the harm-based justification is a common thread running through free speech case law and scholarship. Moreover, that in the absence of a harm-based justification for regulation, the general right to free speech is typically understood as controlling. In other words, subsections 4.1.1–4.1.7 are meant to establish that courts and legal scholars are committed to the proposition: that the general speech right can be abridged if *and only if* there exists a harm-based justification for carving out an exception. To demonstrate this, I provide a capsule tour of First Amendment jurisprudence, highlighting the harm-based argument underlying the justifications for regulation and non-regulation throughout history and across speech categories.

---

[7] *See*, *e.g.*, Becker, *supra* note 6; Steven Shavell, Economic Analysis of Accident Law (Harvard University Press 2007).

[8] Richard Posner, *Free Speech in an Economic Perspective*, 20 Suffolk U. L. Rev. 1 (1986). *See also* Richard Posner, Economic Analysis of Law 955–972 (9th ed. 2014).

[9] *Free Speech*, *supra* note 8, at 3.

When a court indicates that the low social value and high social costs of a class of expressive activity are grounds for allowing regulation of it, we may infer that the court is assuming the harm-based argument diagrammed above. Conversely, when a court points out either the non-negligible value or innocuousness of a class of expressive activity as a ground for protecting it, the court must be assuming that: (1) the harm-based justification is the *only* relevant justification for abridging the general speech right; and (2) that the type of expressions at issue commonly fail to satisfy condition (i) or condition (ii).

In either case, the inferential connection between the relative harms and benefits of the speech category and regulation of that category depends upon the harm-based justification. The reference to harms in the context of speech regulation would not make sense if the validity of harm-based justifications were not an assumed major premise.

4.1.1 General Theories of Speech Rights

Innumerable general theories of speech rights have been offered throughout history. This subsection considers a mere handful of representative samples, highlighting the prevalence of harm-based justifications for allowing abridgment of the general speech right. Unavoidably, the task entails some insensitivity to the nuances of the various positions considered. The reasons given for when and why speech rights should be respected, and when and why they should make way for other concerns are diverse and subtle. However, these subtleties are mostly irrelevant for present purposes. What matters is that the authors ultimately accept some specification of the harm-based argument for abridging speech rights.

It is worth observing first that the notion of a general speech right is not a recent invention. The ancient Athenians claimed to recognize a general right to free speech,[10] although, recalling the death of Socrates, we may well wonder as to its true extent. The nominal right seems in any case to have arisen in various legal systems throughout history. Unfortunately few sustained defenses of the principle are discernible in ancient texts.

John Milton's *Areopagitica* is commonly recognized as being among the earliest modern pleas for liberal governance.[11] However, the reverence in which civil libertar-

[10] *See, e.g.*, Arlene W. Saxonhouse, Free Speech and Democracy in Ancient Athens (2005); Stephen Halliwell, *Comic Satire and Freedom of Speech in Classical Athens*, 111 J. Hellenic Stud. 48 (1991).

[11] John Milton, *Areopagitica*, *in* John Milton: The Major Works 236 (Stephen Orgel & Jonathan Goldberg, eds., 2008).

ians hold the poet is puzzling, for he seemingly would allow considerable incursion upon his posited speech right.[12] The concern which motivated the *Areopagitica* was the licensing and censorship of printed material, and it is only against these particular forms of speech regulation which Milton protested. Milton advocated no limitation whatever on how far governments could punish the authors of blasphemous publications after the fact.[13] In the present-day parlance of law, he opposed merely the "prior restraint" of speech. Indeed, he was apparently an enthusiastic supporter of ex post regulation. And even with respect to prior restraint, Milton's commitment seems tepid relative to present-day standards. He was generous in the exceptions he would carve out of the prohibition on prior restraints of speech—conceding to restrictions on "popery," superstition, impiety, or "evil" speech generally.[14]

Milton's arguments—both for and against speech rights—were predominantly moral arguments. Nevertheless, the *Areopagitica* is a landmark in the evolution of free speech, and it merits at least passing consideration whether even here the harm-based justification may be retrieved. Inasmuch as "popery" and "open superstition" are deemed sufficiently extreme to warrant prior restraint by government intervention, so too are expressions which have the tendency to extirpate "civil supremac[y]."[15] The contention seems to be that those ideas which threaten to undermine the government may justifiably be censored by the government. This at least seems to be an exclusion to the prohibition on prior restraints, founded upon the reduction of a distinctly non-moral harm. Thus do we find—even in the murky moralistic dawn of the modern age—a harm-based justification for the regulation of speech.

Immanuel Kant is also often characterized as a defender of speech rights. Given his association with deontological ethics, it may surprise some readers to learn that he seems at least in some of his work to have endorsed harm-based justifications for abridging the right to free speech. For example, in *What is Enlightenment?*, he writes, "Nothing is required for this enlightenment, however, except freedom, and indeed the most harmless among all the things to which this term can properly be applied. It is the freedom to make public use of one's reason at every point." Here

[12] His arguments *for* speech rights are unimpressive also. Much of his polemic is aimed at establishing the value of speech *for the Christian religion*. Quite a lot of his argument is founded upon appeals to authority and vague references to the "will of God."

[13] *Areopagitica* ("Those which otherwise come forth, if they be found mischievous and libellous, the fire and the executioner will be the timeliest and the most effectuall remedy, that mans prevention can use.")

[14] Id. "I mean not tolerated Popery, and open superstition, which as it extirpats all religions and civill supremacies, so it self should be extirpat, . . . that also which is impious or evil absolutely either against faith or maners no law can possibly permit, that intends not to unlaw it self."

[15] *Id.*

already is cause for doubt. Kant's gratuitous characterization of speech as "harmless" reveals much about his thinking. Why should it *matter* that the freedom of speech is "harmless," unless *harm* were a possible ground for abridgment of the right?

Indeed, Kant seems to have believed precisely this, for he goes on to write that the speech of public officials and pastors might justifiably be abridged on account of the harms which would result from unencumbered expression of their thoughts, owing to the authority of their offices.[16] Harms, therefore, *can* justify the abridgment of speech, even in Kant's relatively liberal view. It merely happens to be the case that—according to Kant—*most* speech simply *cannot be* very harmful.[17] Critically however, this does not foreclose the justifiable regulation of speech if it *were* shown to be harmful.

John Stuart Mill adopts a similar position—though his underlying political and moral theory is quite distinct from Kant's. Mill was arguably most important proponent of speech rights—certainly prior to the twentieth century.

The first two chapters of *On Liberty* express many fine arguments for a liberal speech doctrine. He is generally understood as advancing an extremum position, maximally favoring the right to free speech. Yet even Mill would allow abridgment of the speech right in order to mitigate harms. He is explicit on this point, writing, "[T]he only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others."[18] The principle was not merely a latent possibility for Mill. He furnished concrete examples of the abridgments he had in mind. In the third chapter, he writes:

> [E]ven opinions lose their immunity, when the circumstances in which they are expressed are such as to constitute their expression a positive instigation to some mischievous act. An opinion that corn-dealers are starvers of the poor, or that private property is robbery, ought to be unmolested when simply circulated through the press, but may justly incur punishment when delivered orally to an excited mob assembled before the house of a corn-dealer, or when handed about among the same mob in the form of a placard. Acts of whatever kind, which, without justifiable cause, do harm to others, may be, and in the more important cases absolutely require to be, controlled by the unfavourable sentiments, and, when needful, by the active interference of mankind. The liberty

[16] *Id.*

[17] A Kantian basis for further abridgment of speech rights may be found, for example, in Helga Varden, *A Kantian Conception of Free Speech*, *in* Free Speech in a Diverse World 39 (Deidre Golash ed., 2010).

[18] John Stuart Mill, On Liberty 9 (Hackett Publishing 1978) (1869).

> of the individual must be thus far limited; he must not make himself a nuisance to other people.[19]

Mill's position is clear: the prevention of harms can justify abridgment of the speech right. The example he supplies would fall within the category of "incitement" in present-day American jurisprudence, however the reasoning he employs seems easily extensible to other cognizable exceptions—for example, the publication of state secrets, true threats, and communications facilitating criminal conspiracies. It is unlikely however that Mill would allow much more than this.

Having opened the door to harm-based justifications, Mill's reluctance to admit exceptions for offensive speech, obscenity, false or libelous statements, or otherwise harmful expressions may seem puzzling. On this point, it is worth clarifying that Mill's definition of "harm" is different from mine. Mill's conception is narrower. He does not believe that *any* harm can justify the abridgment of the speech right, because many of the things which satisfy conditions (i) and (ii) would not count as proper "harms" by his lights. He would exclude psychic harms—for example, shock, disgust, loss of dignity. These are not properly "harms" in Mill's theory. For an expression to justify exception to the general speech right, its harm must be both substantial and physical. Mill would surely object to *most* of the speech regulation which liberal democracies presently allow. Yet he would not reject their harm-based justifications on principle.

Mill's insensitivity to psychic harms is somewhat surprising. His utilitarian moral philosophy—equating "pleasure" with the good, and "pain" with bad—seems to ground all ethics upon wholly mental phenomena. He was moreover an economist, making it all the more puzzling that he refrains from drawing the natural equivalence between "harm" with disutility. Possibly he was concerned about the problem of error and the inaccessibility of mental states to third-party observers. Regardless, little interpretative work is required to see the harm-based justification at work in Mill's theory of speech rights. Though not *all* harms—as defined by conditions (i) and (ii)—can justify speech regulation, any acceptable abridgment of the speech right, according to Mill, requires *some* harm-based justification.

Many subsequent moral and political philosophers have accepted some variation of Mill's harm principle. For example, Jeremy Waldron advances a theory substantially similar to Mill's, though expanding the scope of the definition of "harm" to include hate speech.[20] For Waldron, psychic harms are no less "harm" for the purposes of speech rights analysis than physical harms. More generally, philosophers,

[19] *Id.* at 54???.

[20] Jeremy Waldron, The Harm in Hate Speech (Harvard University Press 2012).

psychologists, and lawyers have tended increasingly to recognize psychic and dignitary harms as properly warranting legal redress, and the extension to Mill's harm principle seems a natural and sensible evolution of the theory.

A far more ambitious expansion of Mill's theory may be found in Professor Feinberg's four-volume series, *The Moral Limits of the Criminal Law*. In the first volume, Feinberg articulates a definition of "harm," which is distinct both from my definition and Mill's definition: "[O]nly setbacks of interests that are wrongs, and wrongs that are setbacks to interest, are to count as harms in the appropriate sense."[21] Interests and wrongs are given careful treatment in succeeding sections. The relevant point is that, like Mill, Feinberg's conception is constrained to exclude quite a lot of what we might ordinarily mean by "harm." Unlike Mill, Feinberg's aim was not to provide a general normative theory of *law* nor of justified government interference, but rather more narrowly—as his title indicates—of the *criminal* law. Nevertheless, quite a lot of speech regulation occurs in the form of criminalization, to which Feinberg's theory is directly relevant. Moreover, to the extent that Feinberg's principles are generalizable, it is indirectly informative to non-criminal speech regulations.

Feinberg's harm principle includes things which I would bracket off as "moral." He therefore would reject the strong claim that *only* expressions satisfying conditions (i) and (ii) can justify abridgment of the speech right. However, he does seem to believe that "harms" (as I have defined them) are one of the possible justifications for the abridgment of speech rights. He is therefore at *least* committed to the harm-based justification. Importantly, Feinberg's definition of "harm" does not include the kinds of harms which Waldron would regard as falling within the harm principle—i.e., psychic harms.

Yet the exclusion of psychic harms from the definition of "harm" does not entail that criminal punishment on the basis of psychological effects cannot be justified under Feinberg's theory. The harm principle is not the only operative justification in his framework. Feinberg would also allow abridgment of speech rights on the basis of "offense," the topic of the second volume in his series.[22] Not all offense can justify a criminal sanction. Similar to his narrowing of the meaning of "harm," Feinberg adopts a notion of "serious offense" which requires an element of wrongfulness. For present purposes, the distinction Feinberg draws between "harms" and "offenses," is immaterial. Expressions which cause "harm" or "offense" (by Feinberg's definition) are both "harmful speech" (under my definition).

[21] Joel Feinberg, The Moral Limits of the Criminal Law: I. Harm to Others 36 (1984).

[22] Joel Feinberg, The Moral Limits of the Criminal Law: II. Offense to Others (1985).

The moral component in Feinberg's theory undermines straightforward mapping to the harm-based justification diagrammed above. Nevertheless, it is clear that under Feinberg's theory, no less than his predecessors, the reduction of harms can justify incursions upon the general speech right.

4.1.2 The Structure of First Amendment Review

The foregoing discussion demonstrates the prevalence of harm-based justifications for limiting speech rights from a philosophical perspective. I turn now to the judiciary's treatment of speech rights in the practice of law in the United States.

First Amendment review developed alongside mid-century innovations in Equal Protection and Due Process jurisprudence, recognizing multiple "levels" of constitutional scrutiny.[23] Tiered review for constitutional questions arose in the context of the Fourteenth Amendment in *United States v. Carolene Products Co.*[24] Articulation of the "strict scrutiny" standard followed in cases decided in the following decade.[25] Although the speech right was not directly implicated in these cases, the Court's opinions averred that government regulation of "core freedoms"—including the freedom of speech and freedom of the press—ought to trigger the most stringent standard of review. Accordingly, as First Amendment controversies wended their way to the Supreme Court, government intrusions upon the speech right were assessed within the framework of tiered review.

The nutshell gloss of First Amendment judicial review is that "content-based" regulations warrant strict scrutiny, and "content-neutral" regulations intermediate scrutiny. The rationale is that incursions upon the ideas and beliefs of the citizenry are anathema to the principles of a liberal democracy. Content-based restrictions target semantic content—the very thing feared of an oppressive government—whereas content-neutral restrictions target conduct. Because conduct only incidentally affects the marketplace of ideas, so long as alternative channels of communicating the same information remain available to speakers, its effect was surmised to be less injurious. The distinction trades upon the difference between regulating *what* is said versus *how*, *when*, and *where* it is said. Thus, if the speech affected by a content-neutral regulation can be expressed via alternative channels, it is presumed that the public will not be deprived of access to the content of that speech, and therefore that the expected damage to the free exchange of ideas would be less egregious (and therefore

[23] This correspondence has not been without criticism. For example, *see* Simon & Schuster, Inc. v. Members of New York State Crime Victims Bd., 502 U.S. 105, 125 (1991) (Kennedy, J. dissenting).

[24] 304 U.S. 144, 152 fn. 4 (1938).

[25] Skinner v. State of Oklahoma ex rel. Williamson, 316 U.S. 535, 540 (1942).

less in need of protection). The Fifth Circuit neatly captures the rationale:

> Content-neutral time-place-manner restrictions are examined under intermediate scrutiny, meaning they are permissible so long as they are narrowly tailored to serve a significant governmental interest and leave open ample alternative channels for communication of the information. Content-based time-place-manner restrictions are examined under strict scrutiny, meaning they must be narrowly drawn to effectuate a compelling state interest.[26]

Things are rather more complex in practice of course. For example, even content-based speech regulation is typically scrutinized under a lower standard when the speech is expressed in a government-owned non-public forum—for example, military bases or prisons.[27] Professor Kelso contends that First Amendment review should therefore be understood as consisting in *three* tiers: strict scrutiny, intermediate review, and reasonableness balancing.[28] Others have raised doubts about employing the categorical approach to First Amendment questions at all (and indeed constitutional questions generally). Bunker, Calvert, and Nevin describe a gradual weakening of strict scrutiny over time,[29] blurring the boundaries of tiered review. Professor Shaman recognizes the trend more broadly in Constitutional jurisprudence, and welcomes a dissolution of "rigid" categories of review.[30]

The subtleties of judicial review need not detain us for present purposes. It suffices to observe that even under the courts' most stringent standard, public policy objectives can justify intrusion on speech rights, so long as those objectives are sufficiently "compelling." This has consistently remained the majority position among the justices of the U.S. Supreme Court from the earliest First Amendment cases.

It should be mentioned however that the defeasibility of speech rights has never enjoyed unanimous approval. Justice Hugo Black envisaged a rather more mechanical approach: that once a category of behavior is recognized as speech, the bar on

[26]Sonnier v. Crain, 613 F.3d 436, 441 (2010) (citing Turner Broad. Sys. v. FCC, 520 U.S. 180, 213?14 (1997); Ward v. Rock Against Racism, 491 U.S. 781, 791, (1989); and Perry Educ. Ass'n v. Perry Local Educators' Ass'n, 460 U.S. 37, 46 (1983)).

[27]*See*, e.g., Adderley v. Florida, 385 U.S. 39, 47 (1966); Greer v. Spock, 242 U.S. 828, 836–40 (1976); Brown v. Glines, 444 U.S. 348, 354–58 (1980).

[28]R. Randall Kelso, *The Structure of Modern Free Speech Doctrine: Strict Scrutiny, Intermediate Review, and "Reasonableness" Balancing*, 8 Elon L. Rev. 291 (2016).

[29]Matthew D. Bunker, et al., *Strict in Theory, But Feeble in Fact? First Amendment Strict Scrutiny and The Protection of Speech*, 16 Comm. L. & Policy, 349 (2011).

[30]Jeffrey M. Shaman, *Cracks in the Structure: The Coming Breakdown of the Levels of Scrutiny*, 45 Ohio L.J. 161 (1984).

government intervention should be absolute.[31] Justice Anthony Kennedy, decrying the slipshod manner in which the "compelling state interest" test meandered into First Amendment jurisprudence, expressed similar reservations.[32] Justices William O. Douglas and Antonin Scalia have also been characterized as free speech "absolutists," although a careful consideration of their opinions reveals more complexity than the label suggests.

In the argument framework of the harm-based justification for speech regulation, the "absolutist" position appears to be founded upon a rejection of premise (4). I should add parenthetically that although I do not argue for "absolutism" in this article, I am sympathetic to that viewpoint. *Even if* a speech regulation were shown to be welfare improving, I am skeptical whether it ought to be the government's business to intervene in the realm of ideas, beliefs, and preferences. My purpose in granting premise (4) to proponents of speech regulation is merely to bracket off a digression which might threaten to overwhelm the principal point I hope to advance here: the defectiveness of premise (2). I should reiterate however that I grant (4) *only* for the purposes of argument, and ultimately I believe the harm-based justification falters on *both* premises (2) and (4).

4.1.3 Defamation

It is often remarked how few First Amendment cases were litigated prior to the twentieth century. There are several reasons for the relatively delayed emergence of First Amendment case law. Facially, the language of the First Amendment only limits the power of the *federal* legislature to make laws abridging the freedom of speech.[33] And the United States Congress evinced little interest in regulating speech

[31] *See generally* Hugo L. Black, *The Bill of Rights*, 35 N.Y.U. L. Rev. 865 (1960); *Reflections on Justice Black and Freedom of Speech*, 6 Val. U. L. Rev. 316 (1972); Loren P. Beth, *Mr. Justice Black and the First Amendment: Comments on the Dilemma of Constitutional Interpretation*, 41 J. Pol. 1105 (1979).

[32] Simon & Schuster, Inc. v. Members of New York State Crime Victims Bd., 502 U.S. 105, 126 (1991) ("The inapplicability of the compelling interest test to content-based restrictions on speech is demonstrated by our repeated statement that 'above all else, the First Amendment means that government has no power to restrict expression because of its message, its ideas, its subject matter, or its content.'") (*citing* Police Dept. of City of Chicago v. Mosley, 408 U.S. 92, 95 (1972)).

[33] Curiously, "originalist" proponents of speech regulation have relied upon this distinction to justify incursions into the speech realm. Given that the First Amendment is now understood to apply to states under incorporation (*see* Everson v. Board of Education, 330 U.S. 1 (1947)), the point seems to be that the "intent of the Framers" was only to assign powers in the design of a federal government. Thus, when Madison proclaimed, "The people shall not be deprived or abridged of their right to speak, to write, or to publish their sentiments; and the freedom of

in its early years—with the notable exception of the Alien and Sedition Acts of 1798. But the constitutionality of the Alien and Sediction Acts was never litigated before the Supreme Court, as the power of judicial review was not established until 1803,[34] and the legislation was written to expire in 1801.[35]

Thus, the First Amendment would not have protected individuals from *state* incursions on their speech rights, and the *federal* government seems not to have tried—but for the one exception. Consequently, we have a dearth of First Amendment case law arising from the early decades of the Republic.

However, our historical inquiry need not end here, for the federal constitution is not the sole source of speech rights in American law. Several *state* constitutions also recognized speech rights, and the justifications for the abridgment of those rights are informative. They are informative insofar as they are examples of rationales which have been offered in defense of speech regulation—not because they have any precedential significance.

The early speech cases typically arose from defamation disputes. It is not difficult to retrieve the harm-deterrence argument implicit in the courts' opinions. For example, in *Runkle v. Meyer*,[36] a printer was sued for publishing a scandalous anecdote involving the putative misdeeds of a clergyman. Among the several issues decided in that case was the question whether libel constituted an actionable claim. The court reasoned that because "slander" (i.e., a spoken defamation) was actionable, it should follow that "libel" (i.e., a printed defamation) ought to be actionable also, on the ground that the *harm* caused by written defamation, capable of reaching a wider audience, would likely be greater than the *harm* caused by a spoken one.[37]

The court's logic relies upon the harm-deterrence argument. Leveraging the

---

the press, as one of the great bulwarks of liberty, shall be inviolable," the originalists seem to say, he meant only that it should be inviolable *by the federal government*. 1 Annals of Cong. 434 (1789). State governments, of course, may violate at will. And when George Washington, addressing his soldiers, announced, "[T]he freedom of Speech may be taken away, and, dumb and silent we may be led, like sheep, to the Slaughter," the originalists would contend, he was only talking of the British—not state legislatures. *Founders Online*, National Archives, *available at:* http://founders.archives.gov/documents/Washington/99-01-02-10840. The contention is baffling.

[34]Marbury v. Madison, 5 U.S. 137 (1803).

[35]Later Supreme Court jurisprudence suggests that it would—or at least *should*—have been found unconstitutional. New York Times Co. v. Sullivan, 376 U.S. 254, 576 (1964) ("Although the Sedition Act was never tested in this Court, the attack upon its validity has carried the day in the court of history. Fines levied in its prosecution were repaid by Act of Congress on the ground that it was unconstitutional.").

[36]3 Yeates 518 (1803).

[37]*Id.* at 519 ("The offence of a libel is more heinous, as its circulation of the slander is more extensive . . . .").

harm-of-slander relative to the harm-of-libel in order to establish the actionability of the latter, the court implies that *harmfulness* is the relevant factor in establishing liability. To deploy its argument, the court must assume that slander is prohibited *because* it generates little value and imposes a substantial externality.[38] This implies the general principle that the more harmful an expression, the stronger the justification for regulating it.

The argument advanced by the court is that if the harm of slander is x, and the harm of libel is y, and $y > x$, then it follows that y should be actionable. The implication is that there is some threshold z whether a cause of action should be recognized. And by transitivity, since $x > z$ and $y > x$, it follows that $y > z$. But this relies upon the premise that *harm* is the relevant consideration whether the abrogation of speech rights should be permitted in a civil action. It is only upon this premise that pointing out the relatively greater harmfulness of printed defamation can have any relevance.

In another of the early defamation cases, *Mayrant v. Richardson*,[39] the plaintiff claimed to have lost a congressional election because the defendant had opined to prospective voters, "[Mayrant's] mind was impaired, weakened, and could never be depended on." The court ruled that the words did not constitute slander: First, because the extent of the negative externality was thought to be marginal (the court thought the words more likely to inspire "compassion" than "hatred, ridicule, or contempt").[40] Second, because the expression of an opinion—especially of an individual seeking public office—was at least potentially socially valuable.[41] Thus we find here again, the court's reasoning relies implicitly upon harm-based reasoning: failure to meet conditions (i) and (ii), the utterance was found non-harmful and thus not actionable.

In *Blunt v. Zuntz* an attorney was sued for slander for statements made during

[38] As to the absence of value in the publication, the court mentions that truth would relieve the printer of his liability. *Id.* at 520. As to the negative externality, *see id.* at 519 ("It will not be denied, that if one designedly bespatters another's cloathes with filth, as he passes the street, . . . he would be liable for damages. And shall a printer with his types, blacken the fairest reputation, the choicest jewel we enjoy . . . .").

[39] 1 Nott & McC 347 (1818).

[40] The court stated that for an alleged defamation to be actionable, the externality should either be legal punishment (presumably contemplating false reports to the police or perjury) or measurable injury. *Id.* at 349. The imputation of mental impairment did not, in the view of the court, cause any such injury. *Id.* ("[Mental impairment] was a misfortune and not a fault. It might have been calculated to excite compassion, but not hatred, ridicule or contempt."); and *id.* at 353 ("[T]he words must be of an opprobrious nature, and such as are calculated to lessen the person of whom they are spoken, in the opinion of the community.").

[41] *Id.* at 353.

litigation.[42] The court notes:

> It is an established rule, that no action can be maintained against a counsellor for words pertinent to the issue, spoken in judicial proceedings . . . . The object of this rule is closely connected with the *utility of the function of counsel*, which consists principally in a liberal freedom of speech, and that he may not be embarrassed by continually balancing in his mind whether the remark he is about to make be slanderous or not.[43]

The court indicates that the relevant factor was failure of condition (i)—that the utterances of lawyers in courts are not of low social value. Though the speech may impose some negative externality, the court recognized the value of it to be *non-negligible*—i.e., that there exists "utility" in "the function of counsel." Ergo, such speech is not harmful, and its expression should not be actionable.

Harm-based reasoning is thus clearly observable in the early speech cases of American courts. Defamation claims would not be tested against the federal constitution until the mid-twentieth century, however the subsequent cases would fundamentally track the harm-based concerns expressed in *Runkle*, *Mayrant*, and *Blunt*.

In its first major defamation case, *New York Times Co. v. Sullivan*,[44] the U.S. Supreme Court ratcheted up protection of speech rights against defamation claims, holding that plaintiffs were required to prove "actual malice" to recover. Critically, like in *Blunt*, the Court's reasoning in *Sullivan* was founded on the failure of condition (i), identifying a non-negligible social benefit in potentially defamatory expressions.[45]

Later, in a distinguishing case limiting the *Sullivan* holding to public figures, the Court in *Gertz v. Robert Welch, Inc.* again weighed the balance of benefits and costs, with Justice Powell writing, "The legitimate state interest underlying the law of libel is the compensation of individuals for the *harm* inflicted on them by defamatory falsehood."[46] And in *Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc.*, retreating further from *Sullivan*, Powell opined, "In light of the reduced constitutional *value of*

[42] Ant.N.P. (2d Ed.) 246.

[43] *Id.* at fn 1 (emphasis added).

[44] *Supra* note 35.

[45] Sullivan, *supra* note 35 at 281 ("The importance to the state and to society of such discussions is so vast, and the advantages derived are so great that they more than counterbalance the inconvenience of private persons whose conduct may be involved, and occasional injury to the reputations of individuals must yield to the public welfare, although at times such injury may be great. The public benefit from publicity is so great and the chance of injury to private character so small that such discussion must be privileged.") (*quoting* Coleman v. MacLennan, 78 Kan. 711, 724 (1908)).

[46] 418 U.S. 323, 341 (1974) (emphasis added).

speech involving no matters of public concern, we hold that the state interest adequately supports awards of presumed and punitive damages—even absent a showing of 'actual malice.'"[47]

It is abundantly clear from the Court's language—both when strengthening and weakening the speech right—that it *is* attempting to identify whether and to what extent conditions (i) and (ii) apply. The harm-deterrence argument is implied throughout the opinions. The Court consistently reasons in terms of the deterrence objective—i.e., decreasing the supply of harmful speech—which defamation actions are meant to effect.[48]

Notably, the Court has been explicit that when abridging the speech right, it is *not* recognizing a latent definitional quirk in the Framers' conception of "speech." The opinions do not ground the defamatory speech exclusion on the basis of a "historical understanding" of the speech right. They apply a balancing test essentially identical to the implicit harm-based rationale employed in the nineteenth century in order to determine whether a category of expression warrants protection.

4.1.4 Espionage Act Cases and Incitement

The first significant speech cases properly litigated under the First Amendment arose in connection with the Espionage Act of 1917 and Sedition Act of 1918. In *Schenck v. United States*,[49] the defendant Charles Schenck, in his capacity as General Secretary of the Socialist Party, oversaw the printing and distribution of leaflets encouraging draftees to resist conscription. The government charged that he had violated provisions of the Espionage Act, which prohibited efforts "to obstruct the recruiting and enlistment service of the United States, when the United States was at war with the German Empire."[50] Writing for a unanimous Court, Justice Holmes devised his

[47] 472 U.S. 749, 760 (1985).

[48] *See, e.g.*, Sullivan, *supra* note 35, at 279 ("Allowance of the defense of truth, with the burden of proving it on the defendant, does not mean that only false speech will be deterred."); *id.* at 299 ("Such criticism cannot [presumably "should not"], in my opinion, be muzzled or deterred by the courts at the instance of public officials under the label of libel."); Gertz, *supra* note 46, at 350 ("[Punitive damages for libel] are not compensation for injury. Instead, they are private fines levied by civil juries to punish reprehensible conduct and to deter its future occurrence."); Dun, *supra* note 47, at 762 ("[I]n the libel context, the States' regulatory interest in protecting reputation is served by rules permitting recovery for actual compensatory damages upon a showing of fault. Any further interest in deterring potential defamation through case-by-case judicial imposition of presumed and punitive damages awards on less than a showing of actual malice simply exacts too high a toll on First Amendment values.") (White, J., concurring in judgment).

[49] 249 U.S. 47 (1919).

[50] *Id.* at 49.

infamous "clear and present danger" test:

> The most stringent protection of free speech would not protect a man in falsely shouting fire in a theatre and causing a panic. It does not even protect a man from an injunction against uttering words that may have all the effect of force. The question in every case is whether the words used are used in such circumstances and are of such a nature as to create a clear and present danger that they will bring about the substantive evils that Congress has a right to prevent. It is a question of proximity and degree. When a nation is at war many things that might be said in time of peace are such a hindrance to its effort that their utterance will not be endured so long as men fight and that no Court could regard them as protected by any constitutional right. It seems to be admitted that if an actual obstruction of the recruiting service were proved, liability for words that produced that effect might be enforced.[51]

Holmes's opinion indicates a balancing of utilities: the Socialist's Party's interest in expression, the public's interest in the receipt of ideas, and the government's interest in advancing the war effort. He evidently felt the war effort to be so weighty an interest as to render all other considerations negligible. He accordingly found the regulations constitutional.

Holmes was careful to distinguish that in different circumstances, the calculus of costs and benefits could change. The implication is that the harmfulness of speech—i.e., whether it satisfies conditions (i) and (ii)—is context-dependent, and therefore that the constitutional protection of speech ought to be correspondingly context-dependent.

It is of critical importance to recognize that this conclusion follows only if Holmes assumes the major premise: "If harmful speech is punished, then social welfare is improved." There is no conceivable other purpose to his pointing out the harmfulness of Schenck's leafletting except to establish the *reduction of harm* as a ground for abridging the speech right.

This framework is reinforced in *Abrams v. United States*,[52] the facts of which are substantially similar to those in *Schenck*, where Holmes wrote in dissent:

> I do not doubt for a moment that by the same reasoning that would justify punishing persuasion to murder, the United States constitutionally may punish speech that produces or is intended to produce a clear and

[51] *Id.* at 52.

[52] 250 U.S. 616 (1919).

> imminent danger that it will bring about forthwith certain substantive evils that the United States constitutionally may seek to prevent. The power undoubtedly is greater in time of war than in time of peace because war opens dangers that do not exist at other times.
>
> But as against dangers peculiar to war, as against others, the principle of the right to free speech is always the same. It is only the present danger of immediate evil or an intent to bring it about that warrants Congress in setting a limit to the expression of opinion where private rights are not concerned. Congress certainly cannot forbid all effort to change the mind of the country. Now nobody can suppose that the surreptitious publishing of a silly leaflet by an unknown man, without more, would present any immediate danger that its opinions would hinder the success of the government arms or have any appreciable tendency to do so.[53]

History of course has noted Holmes's apparently shifting disposition. In *Schenck*, he found the wartime leafletting to have satisfied the requirements of the "clear and present danger" test, while in *Abrams*, under substantially the same set of facts, he found that it did not. The reason for the apparent transformation has fueled much speculation among historians and biographers.[54] Yet Holmes contended his position had not changed.[55] Historians propose two possible assessments: first, that Holmes inexplicably failed to notice that he had arrived at divergent outcomes when presented with substantially identical facts; and second, that his protestations of self-consistency were intellectually dishonest.

I believe there is a third possibility. What Holmes may have meant when he claimed his position had not changed was that his reliance on the harm-based argument remained consistent. In his dissent, Holmes reaffirms the premise that the harmfulness of speech *can* justify carving out exceptions to the First Amendment. The distinction is merely that in *Abrams* he no longer seemed to regard anti-war leafletting as satisfying the definition of "harmful speech." What altered his assessment is unknown. Possibly he thought Schenck's leaflets more harmful than Abrams's leaflets, or possibly he reconsidered his assumptions about the harm of leafletting generally. In either case, it would have been the minor premise—not the rule—which determined the different outcomes.

[53] *Id.* at 627–628 (Holmes, J. dissenting).

[54] Holmes, writing for the majority, reaffirmed the "clear and present danger" test in two other pre-*Abrams* cases, Frohwerk v. United States, 249 U.S. 204 (1919) and Debs v. United States, 249 U.S. 211 (1919). His use of "clear and *imminent* danger" rather than "clear and present danger" in *Abrams* is also curious.

[55] *Id.* at 627.

Viewed in the context of Holmes's approach to law generally, it is unsurprising that his First Amendment jurisprudence relied upon distinctly economic premises. It was his sensitivity to the economic tradeoffs in rule-making—more than any particular set of decisions—which has proved to be his greatest contribution to the law.

The economic torch was duly passed in *United States v. Dennis*, in which the succeeding generation's great proto-economist, Judge Learned Hand, elaborating upon Holmes's "clear and present danger" test, offered a yet more explicitly economic construction of the rule:

> In each case [courts] must ask whether the gravity of the 'evil,' discounted by its improbability, justifies such invasion of free speech as is necessary to avoid the danger.[56]

Hand's analysis is uncannily familiar, for it closely resembles the negligence formula he famously articulated in *United States v. Carroll Towing Co.*.[57] Hand's opinion in *Carroll Towing*, ubiquitous in torts casebooks, prescribes a balancing of expected benefits against expected costs to effect efficient precautionary care incentives for prospective injurers. The *Dennis* elaboration of the "clear and present danger" test simply extends the tort principle to the First Amendment context.

Interestingly, prior to *Dennis*, when Hand was still a district court judge, he had fashioned a rather different First Amendment test in *Masses Publishing Co. v. Patten*.[58] The plaintiff—publisher of the socialist magazine *The Masses*—moved for preliminary injunction against the postmaster, who refused to deliver the publication on the ground that it was "seditious" under the terms of the Espionage Act. The magazine was critical of the decision to go to war and praised citizens who would resist conscription. Hand's test in *Masses* was to require that the incitement be *direct* to justify exception to the general speech right.[59] Thus, legislation prohibiting the expression, "You must violently overthrow the government," would be constitutional, whereas prohibition of the expression, "It would be laudable if someone

[56] United States v. Dennis, 183 F.2d 201, 212 (2d Cir. 1950), *aff'd*, 341 U.S. 494 (1951).

[57] 159 F.2d 169 (2d. Cir. 1947) (defining the duty of care as $B < PL$, or the burden of precautions less than the probability of loss). Posner, of course remarks upon the convergence of principle. *Free Speech*, *supra* note 8 at 8; *Economic Analysis of Law*, *supra* note 8 at 958.

[58] 244 F. 535 (1917).

[59] *Id.* at 540 ("[T]o assimilate agitation, legitimate as such, with direct incitement to violent resistance, is to disregard the tolerance of all methods of political agitation which in normal times is a safeguard of free government. The distinction is not a scholastic subterfuge, but a hard-bought acquisition in the fight for freedom, and the purpose to disregard it must be evident when the power exists. If one stops short of urging upon others that it is their duty or their interest to resist the law, it seems to me one should not be held to have attempted to cause its violation.").

would overthrow the government," would not be. The distinction is a crystalline one, for it reduces the factual question to one of linguistics. The former expression is imperative, whereas the latter is declarative.

Throughout his *Masses* opinion, Hand repeatedly emphasized the high social value of political dissent as a ground for narrowly construing the Espionage Act. Thus, even before the question went to the Supreme Court in *Schenck*, it seems Hand had already envisaged that conditions (i) and (ii) would need to be satisfied to overcome the presumption favoring the freedom of expression. Though the *Masses* test is distinct from the "clear and present danger" test, a convergence in reasoning is observable. Both approaches seek to formulate a harm-based framework for limiting the general speech right.[60]

4.1.5 Fighting Words and Hate Speech

A third context in which the courts have recognized limitations to the general speech right is the "fighting words" exception. The precedent was established in *Chaplinsky v. New Hampshire*, in which Justice Murphy, writing for a unanimous Court, premised the constitutionality of speech regulations upon whether the targeted expressions, "by their very utterance inflict[ed] injury or tend[ed] to incite an immediate breach of the peace."[61]

In the facts which gave rise to the controversy, a proselytizing Jehovah's Witness was fined for insulting a police officer, calling him a "damned racketeer," and a "damned Fascist," in violation of New Hampshire laws prohibiting expressions susceptible to provoking violence. In justifying the holding, Murphy explained:

> It has been well observed that such utterances are no essential part of any

[60] In *Masses*, Hand would grant the plaintiff's motion for preliminary injunction on the ground that condition (i) was not satisfied. The Second Circuit reversed. Masses Pub. Co. v. Patten, 244 F. 535 (1917). And the Supreme Court, as we have seen, would ultimately chose a different rule—at least initially. The story does not quite end here, however, for Hand's test is widely regarded as having inspired the "imminent lawless action" test in *Brandenburg v. Ohio*, which would later displace the "clear and present danger" test. 395 US 444 (1969).

Although *Brandenburg* remains the standard for evaluating the constitutionality of incitement laws today, the published opinions are not particularly enlightening for present purposes. Chief Justice Warren's majority opinion gives no reason for the departure from the Holmes cases, pretending continuity with *Dennis*. Justice Black's concurring opinion, consisting of one paragraph, merely indicates his view that *Brandenburg* does invalidate the "clear and present danger" test. And Justice Douglas's concurrence takes the position that the First Amendment should not be subject to *any* exceptions, and thus supplies no reasoning to justify why the Court carves out the limits that it does.

[61] 315 U.S. 568, 572 (1942).

> exposition of ideas, and are of such *slight social value* as a step to truth that any *benefit* that may be derived from them is *clearly outweighed* by the social interest in order and morality.[62]

The identification of conditions (i) and (ii) could not more clearly have been within the contemplation of the Court. Moreover, the harm-deterrence argument is also evident in the Court's reasoning. Envisioning a deterrent effect leading to an improvement in social welfare, Murphy wrote:

> [T]he statute had been previously construed as intended to preserve the public peace by punishing conduct, the direct tendency of which was to provoke the person against whom it was directed to acts of violence.[63]

The language in *Chaplinsky* does not establish the stronger proposition—implied in some later cases—that harmfulness is the *only* justification for the abridgment of the general speech right. Murphy writes about balancing "social value" against the "social interest in order *and morality*." This is ambiguous, potentially leaving room for non-consequentialist considerations. Regardless, the Court in *Chaplinsky* was indubitably asserting that social cost can be *one* justification for abridging speech rights.

Chaplinsky would provide the foundation for the Court's hate speech doctrine. These cases are an interesting avenue for further exploration. In *R.A.V. v. City of St. Paul*,[64] a divided Court—though unanimous in judgment—ruled upon the constitutionality of a municipal ordinance proscribing expressive activities which might arouse "anger, alarm or resentment in others on the basis of race, color, creed, religion or gender."[65]

The petitioner was charged under the St. Paul ordinance for burning a cross in the front yard of a Black family. Justice Scalia, writing for the majority, premised his judgment upon the legislation's use of an "impermissible motive." The gravamen of his argument was that even where legislation falls within a recognized exception to the general speech right, prohibiting a *subset* of that excepted class requires an independent justification. Absent this, Scalia's concern was evidently that *Chaplinsky* could be used as a workaround to allow content-based regulations:

> That would mean that a city council could enact an ordinance prohibiting only those legally obscene works that contain criticism of the city

[62] *Id.* (emphasis added).
[63] *Id.* at 574, fn 8.
[64] 505 U.S. 377 (1992).
[65] *Id.* at 377.

> government or, indeed, that do not include endorsement of the city government. . . . The proposition that a particular instance of speech can be proscribable on the basis of one feature (e.g., obscenity) but not on the basis of another (e.g., opposition to the city government) is commonplace and has found application in many contexts. We have long held, for example, that nonverbal expressive activity can be banned because of the action it entails, but not because of the ideas it expresses—so that burning a flag in violation of an ordinance against outdoor fires could be punishable, whereas burning a flag in violation of an ordinance against dishonoring the flag is not.[66]

Scalia's apprehension seems to be a sensible one, which the hostile concurring opinions may not have fully appreciated. However, the doctrinal solution he devised seems to have been confused—or at least profoundly unsatisfying.

Scalia apparently understood *Chaplinsky* to do no more than carve out a constitutional exception for the category "fighting words." Yet *Chaplinsky* does more than this. It establishes a cost-benefit framework for evaluating First Amendment controversies. Having determined—sensibly—that the abrogation of speech rights within the subset of a recognized exception requires independent justification, Scalia does not proceed to the natural followup question whether the expressions proscribed under the St. Paul ordinance would satisfy the *Chaplinsky* test. Rather, he simply strikes down the ordinance on the ground that it is content-based.

The concurring opinions of Justices White, Blackmun, and Stevens are unsparing in their criticism of the majority. White lambasts Scalia's proposed requirement that the regulation of a subset of an already recognized First Amendment exception must have an independent content-neutral justification. White contends that speech within a recognized exception may permissibly be proscribed without further inquiry.[67]

Blackmun's brief concurrence is ambiguous, on the one hand reasoning that if the regulation of fighting words is permissible, then its more harmful variants ought—on the basis of being more harmful—be regulable *a fortiori*. This echoes the argumentative move the *Runkle* court used in recognizing the actionability of libel claims.[68] Yet on the other hand, Blackmun writes also that the parsing of recognized exceptions

[66] *Id.* at 384.

[67] *Id.* at 400 ("To the contrary, those statements meant precisely what they said: The categorical approach is a firmly entrenched part of our First Amendment jurisprudence. . . . [T]he majority holds that the First Amendment protects those narrow categories of expression long held to be undeserving of First Amendment protection—at least to the extent that lawmakers may not regulate some fighting words more strictly than others because of their content.").

[68] *Supra* note 36.

could lead to a dissolution of speech *protections*. Assuming he did not think the erosion of speech rights a desirable outcome, his position seems puzzlingly inconsistent.

Stevens's concurrence seems to provide the widest perspective.[69] Stevens recognized that Scalia's position, embracing a set of speech *categories* as exceptions to the First Amendment, is at odds with the Court's treatment of speech categories. Scalia would allow only a categorical—not a content-based—incursion upon the general speech right. Stevens contends however that *all* the Court's First Amendment exceptions are fundamentally grounded upon the harm-based justification following *Chaplinsky*. The categorical exceptions are exceptions *because* they are harmful—which is very often a *content-based* determination. Thus, the ground for a First Amendment exception according to Stevens should not be mere membership within a pre-established category, but rather the expected harmfulness of expressions typical of the category. If the representative expression within the class satisfies conditions (i) and (ii), then that class of expressions warrants less protection.

R.A.V. is a confused case. For present purposes, its most notable feature is that Scalia's opinion eschews harm-based arguments entirely. His opinion proposes a purely legalistic test: x is a permissible regulation of speech if and only if x is neither more nor less than the whole category of obscenity, fighting words, criminal conspiracy, defamation, etc. The *reasons why* these categories have been carved out of the general speech right are given no serious consideration. Scalia declines to say whether further categories can be added to this list, or whether changing circumstances might warrant removing a category.[70] The basis is characterized in formalistic terms (patterned on "legal logic" rather than consideration of consequential effects).

The concurrences, though not fully appreciative of the concern motivating Scalia's opinion, are correct that Scalia's formalistic approach is at odds with the Court's prior treatment. Pointing out the harm-based ground for the St. Paul ordinance, White insisted:

> This selective regulation reflects the city's judgment that harms based on race, color, creed, religion, or gender are more pressing public concerns than the harms caused by other fighting words. In light of our Nation's long and painful experience with discrimination, this determination is plainly reasonable. Indeed, as the majority concedes, the interest is compelling.[71]

[69] 505 U.S. at 417.

[70] Scalia acknowledges that "compelling state interests" may override the general speech right, conceding to cost-benefit weighing in principle, though he seems to regard the St. Paul ordinance as overly broad in achieving the government's objective. *Id.* at 395–96.

[71] *Id.* at 407.

Blackmun's concurrence also mentions the relative harmfulness of generic fighting words as compared with racially charged symbols.[72] And Stevens discusses at length the harmfulness justification in limiting First Amendment rights in several contexts.[73]

Scalia's opinion is interesting in the present context because it represents one of the few genuine exceptions where the Court has employed non-harm-based justifications for speech regulation. It is important to recognize however that in *R.A.V.*, the result was not to *allow* the disputed regulation, but rather to invalidate it. Whether Scalia would have considered harmfulness when deciding to permit an exception is unclear.

Discerning a consistent thread in Scalia's First Amendment jurisprudence is a difficulty which has motivated much commentary.[74] Regardless, taken in the context of succeeding cases, *R.A.V.* seems not to represent a general departure from harm-based justifications for speech regulation for the Court as a whole.

In a factually similar subsequent case, *Virginia v. Black*,[75] the Court struggled to apply Scalia's content/category test from *R.A.V.* Justice O'Connor, writing for the majority in *Black*, distinguished that the Virginia statute at issue, which prohibited cross-burning specifically, was constitutional under the "true threats" exception—though the statutory specification that cross-burning should be taken as prima facie evidence of intimidation was not constitutional (being "content-based"). As Justice

[72] *Id.* at 415.

[73] *Id.* at 423, 424–26, 429–435.

[74] Justice Scalia's First Amendment jurisprudence is a fascinating web of apparently contradictory positions, which I cannot satisfactorily explore within the scope of this article. On the one hand, he established a reputation as a stalwart defender of individual speech rights in cases like *R.A.V.*, id.; Texas v. Johnson, 491 U.S. 397 (1989) (finding prohibitions on flag-burning unconstitutional); and Brown v. Entertainment Merchants Association, 564 U.S. 786 (2011) (striking down a California law banning the sale of violent video games to children). On the other hand, he voted to allow the regulation of speech in cases like Hazelwood School District v. Kuhlmeier, 484 U.S. 260 (1988) (carving out an exception for student speech in a curricular context); Morse v. Frederick, 551 U.S. 393 (2007) (expanding the scope of "school speech" to school sanctioned events generally).

Unlike his judicial philosophy in other contexts, Scalia's First Amendment decisions have only vaguely implicated "originalism." *See* discussion, *supra* note 5, 33. *See also* Antonin Scalia, *Originalism: The Lesser Evil*, 57 U. Cinn. L. Rev. 849 (1989). There are traces of his "originalist" philosophy to be found in *R.A.V.* and also in his dissent in Board of County Com'rs, Wabaunsee County, Kan. v. Umbehr, 518 U.S. 668, 686 (1996) (curiously characterizing the majority's position as "formalist"). On Scalia's First Amendment jurisprudence generally, *see* David Schultz, *Justice Antonin Scalia's First Amendment Jurisprudence: Free Speech, Press and Association Decisions*, 9 J. L. & Pol. 515 (1993); Jay S. Bybee, *Common Ground: Robert Jackson, Antonin Scalia, and a Power Theory of the First Amendment*, 75 Tul. L. Rev. 251 (2000); and the symposium articles in 15 First Amend. L. Rev. 152–330 (2017).

[75] 538 U.S. 343 (2003).

Souter observes in his dissent, distinguishing the prima facie evidentiary provision from the substantive prohibition cannot sensibly save the statute, for the prohibition itself (not merely the presumption clause) specifies *cross-burning* rather than "true threats" generally.[76] O'Connor's response was to distinguish Virginia's statute—singling out cross-burning—as consistent with *R.A.V.*, because cross-burning is an especially virulent form of intimidation.[77] It constituted a permissible partition of a recognized exception, because the *reason* for regulating the subset was identical to the *reason* for regulating the category generally. The distinction seems a flaccid one, for one might just as well say that the St. Paul ordinance was directed at a subset of "fighting words," which are especially virulent in the form of a burning cross.

Regardless how the cross-burning cases should be reconciled, or how the Court would decide future cross-burning cases, the relevant point is that to the extent that *R.A.V.* seemed to eschew the harm-based justification, *Black* grafted it onto the content/category test in allowing "particularly virulent" subsets to be regulable independently of a category. Thus, even if *R.A.V.* were understood to represent an abandonment of harm-based reasoning, the harm-based framework was reinstated in *Black*.

One last representative case worth considering in the context of fighting words and hate speech is *Snyder v. Phelps*,[78] where members of the infamous Westboro Baptist Church picketed at the funeral of a fallen marine with signs reading, "God hates fags," and "Fags doom nations."[79] They were sued by the soldier's family for intentional infliction of emotional distress. Writing for the majority, Justice Roberts recognized the distasteful speech to be constitutionally protected, emphasizing the

[76] VA Code Ann. § 18.2–423 states:

> It shall be unlawful for any person or persons, with the intent of intimidating any person or group of persons, to burn, or cause to be burned, a cross on the property of another, a highway or other public place. Any person who shall violate any provision of this section shall be guilty of a Class 6 felony.
>
> Any such burning of a cross shall be prima facie evidence of an intent to intimidate a person or group of persons.

[77] *Id.* at 363 ("It shall be unlawful for any person or persons, with the intent of intimidating any person or group of persons, to burn, or cause to be burned, a cross on the property of another, a highway or other public place. Any person who shall violate any provision of this section shall be guilty of a Class 6 felony. Any such burning of a cross shall be prima facie evidence of an intent to intimidate a person or group of persons.").

[78] 562 U.S. 443 (2011).

[79] *Id.* at 469. It is perhaps worth mentioning, given the theme of the signage, that the deceased was not gay. Evidently, members of the Westboro Baptist Church inferred his death to have been vengeance for what they perceived to be a generalized moral deterioration of Christian values.

non-negligible value of political speech:

> The "content" of Westboro's signs plainly relates to public, rather than private, matters. The placards highlighted issues of public import—the political and moral conduct of the United States and its citizens, the fate of the Nation, homosexuality in the military, and scandals involving the Catholic clergy—and Westboro conveyed its views on those issues in a manner designed to reach as broad a public audience as possible. Even if a few of the signs were viewed as containing messages related to a particular individual, that would not change the fact that the dominant theme of Westboro's demonstration spoke to broader public issues.[80]

Restated in the framework of harmfulness, Roberts's opinion indicates the failure of condition (i). Because the speech is not of negligible value, the speech is not "harmful speech," therefore the harm-based justification fails, and therefore there is no exception to the general speech right protecting it.

Justice Alito, the lone dissenter, engaged the majority within the framework of harmfulness, citing *Chaplinsky* and contending condition (i) was satisfied: "Allowing family members to have a few hours of peace without harassment does not undermine public debate";[81] and that condition (ii) was also satisfied: that the picketing would have the "potential to wound as a personal verbal assault on a vulnerable private figure."[82]

the distinction between speech relating to matters of public versus private concern. The doctrine is, I think, little more than a categorical presumption relating to the conditions of harmful speech. Pulling on the thread, it reveals a neat alignment with the harm-based justification, relying on the additional premise that matters of public concern *presumptively* fail condition (i). The argument for the presumption is that matters relating to politics and social institutions are so paramount in importance, that any limitation of *that* type of speech should be especially disfavored.[83]

[80] *Id.* at 444.

[81] *Id.* at 473.

[82] *Id.* at 475.

[83] The Court's unwavering commitment to political speech over other categories seems myopic. I would hazard to suppose that if violinists were polled what kinds of expression deserved the most stringent protection, they might name music; and if painters, then paintings; and if physicists, then physics publications. The privileging of political speech seems to be founded upon little more than the vanity of political actors. Looking across the broad sweep of history, I can think of no statesman, activist, revolutionary, or judge whose expressive contributions compare favorably to those of Bach, Shakespeare, Picasso, or Newton. And I see no sensible reason why political speech should be privileged over expressions in art, science, mathematics, or philosophy. Presumably,

Put differently, the public/private distinction is premised upon the proposition that political speech (broadly construed) is *so* important that it should be singled out as failing condition (i) by default.[84]

4.1.6 Commercial Speech

Another realm of expression where the Court has commonly allowed regulation is "commercial speech." The rationale tracks the "public versus private concern" distinction previously discussed, which in turn hinges upon the satisfaction of condition (i).

As a historical matter, the Courts' development of its commercial speech doctrine was erratic. In *Valentine v. Chrestensen*, the Court held, "We are . . . clear that the Constitution imposes no . . . restraint on government as respects purely commercial advertising."[85] The Court cites no precedents, nor provides any argument for the

the thought is that political speech is most important, because it is essential to a well-functioning liberal democracy. And poorly functioning social institutions—especially those of an autocratic or oppressive government—might impinge upon all other areas of speech. Thus, if any category of speech is valuable, then political speech must be the most valuable of all. *See* cases cited *supra* note 84. This argument is weak. Even if the premise were conceded, the form of the argument is to privilege means over ends. It is like claiming that instruments are more valuable than music, because without instruments there would be no music. How unquestioningly courts seem to accept this plainly fallacious reasoning is nothing short of baffling. Yet regardless whether this is right, the point is that the Court has viewed political speech as deserving special protection *because* it regards expressions of political speech as being distinctly unlikely to be expressions which satisfy condition (i).

[84] *See*, *e.g.*, Monitor Patriot Co. v. Roy, 401 U.S. 265, 272 (1971) ("[I]t can hardly be doubted that the constitutional guarantee has its fullest and most urgent application precisely to the conduct of campaigns for political office."); Whitney v. California, 274 U.S. 357, 375 (1927) (Brandeis, J. concurring) ("Those who won our independence believed that the final end of the state was to make men free to develop their faculties, and that in its government the deliberative forces should prevail over the arbitrary. . . . They believed that freedom to think as you will and to speak as you think are means indispensable to the discovery and spread of political truth; that without free speech and assembly discussion would be futile; that with them, discussion affords ordinarily adequate protection against the dissemination of noxious doctrine; that the greatest menace to freedom is an inert people; that public discussion is a political duty; and that this should be a fundamental principle of the American government."); Mills v. State of Alabama, 384 U.S. 214, 218–219 ("Whatever differences may exist about interpretations of the First Amendment, there is practically universal agreement that a major purpose of that Amendment was to protect the free discussion of governmental affairs. This of course includes discussions of candidates, structures and forms of government, the manner in which government is operated or should be operated, and all such matters relating to political processes.").

[85] 316 U.S. 52, 54 (1942).

principle. Subsequent scholarship and judicial opinions have regarded the decision unfavorably.[86]

In truth, the doctrine was hardly "clear" in the common law of the states prior to *Valentine*. For example, in *People v. Osborne*,[87] a California court ruled that the state could not permissibly interfere in the speech of a barber, who was fined for displaying prices for his services outside his shop in violation of a city ordinance.[88] The court observed:

> The constitutional liberty of speech implies the right to freely utter and publish whatever the citizen may please and immunity from legal censure and punishment for the publication so long as it is not harmful in its character when tested by such standards as the law affords, and what may be spoken may be written.[89]

Similarly, commercial speech in the context of newspaper advertising was recognized in earlier cases as protected speech.[90] Yet regardless, *Valentine* was the first direct treatment of "commercial speech" by the U.S. Supreme Court in relation to the federal constitution. And the Court proclaimed no protection whatever of its expression.

Doubts about the doctrine grew steadily in the ensuing decades.[91] *Valentine* was finally overturned in *Virginia State Pharmacy Board v. Virginia Citizens Consumer Council*,[92] which concerned a statute declaring it unprofessional for pharmacists

[86] *See, e.g.*, Virginia State Pharmacy Board v. Virginia Citizens Consumer Council, 425 U.S. 748 (1976); 44 Liquormart, Inc. v. Rhode Island, 517 U.S. 484, 522 (1996) (Thomas, J., concurring in judgment); Alex Kozinski & Stuart Banner, *Who's Afraid of Commercial Speech?*, 76 VA. L. Rev. 627 (1990).

[87] 59 P.2d 1083 (1936). *See also* Alex Kozinski & Stuart Banner, *The Anti-History and Pre-History of Commercial Speech*, 71 Tex. L. Rev. 747, 763–772 (1993) (describing several similar cases and the history preceding *Valentine* in considerable detail).

[88] Predating incorporation of the First Amendment (*See Carolene Products*, 304 U.S. 144 (1938) and *Everson*, 330 U.S. 1 (1947)), *Osborne* was decided in the context of the speech guarantee of the California Constitution.

[89] 59 P.2d at 1087 (*citations omitted*).

[90] *See, e.g.*, Ex parte Jackson, 96 U.S. 727, 733 (1878); In re Rapier, 143 U.S. 110, 134 (1892); Lewis Publishing Co. v. Morgan, 229 U.S. 288, 315 (1913).

[91] *See, e.g.*, Cammarano v. United States, 358 U.S. 498, 514 (1959) (Douglas J., concurring); Pittsburgh Press Co. v. Pittsburgh Comm'n on Human Relations, 413 U.S. 376, 401, fn. 6 (1973) (Stewart, J., dissenting); *id.* at 404 (Blackmun, J., dissenting); Lehman v. City of Shaker Heights, 418 U.S. 298, 314–315 fn.6 (1974) (Brennan, J., dissenting); Bigelow v. Virginia, 421 U.S. 809 (1975).

[92] 425 U.S. 748 (1976).

to advertise prices, fees, premiums, discounts, rebates, or credit terms.[93] Justice Blackmun, writing for the majority, rejected the proposition that commercial speech presumptively satisfied the conditions of harmful speech:

> Our question is whether speech which does no more than propose a commercial transaction is so removed from any exposition of ideas, and from truth, science, morality, and arts in general, in its diffusion of liberal sentiments on the administration of Government, that it lacks all protection. Our answer is that it is not.[94]

It is revealing to observe the *way* that Blackmun attacks the commercial speech exception established in *Valentine*. He disputes whether commercial speech is presumptively harmful speech, explaining how plausible it would be for instances of commercial speech to fail conditions (i) and (ii). Describing why it ought not be assumed that commercial speech satisfies condition (i):

> So long as we preserve a predominantly free enterprise economy, the allocation of our resources in large measure will be made through numerous private economic decisions. It is a matter of public interest that those decisions, in the aggregate, be intelligent and well informed. To this end, the free flow of commercial information is indispensable. . . . even if the First Amendment were thought to be primarily an instrument to enlighten public decisionmaking in a democracy, we could not say that the free flow of information does not serve that goal.[95]

And why it ought not be assumed that commercial speech satisfies condition (ii):

> Price advertising, it is said, will reduce the pharmacist's status to that of a mere retailer. The strength of these proffered justifications is greatly undermined by the fact that high professional standards, to a substantial extent, are guaranteed by the close regulation to which pharmacists in Virginia are subject. And this case concerns the retail sale by the pharmacist more than it does his professional standards. Surely, any pharmacist guilty of professional dereliction that actually endangers his customer will promptly lose his license.

Subsequent courts have repeated in unequivocal terms the proposition that the justification for regulating commercial speech is harm-based. For example, in *City of*

[93] Va. Code Ann. § 54–524.2(a) (1974).
[94] 425 U.S. at 762 (*citations omitted*).
[95] *Id.* at 765.

Cincinnati v. Discovery Network, Inc.,[96] the Supreme Court wrote, "[The government] has not asserted *an interest in preventing commercial harms* by regulating the information distributed by respondent publishers' newsracks, which is, of course, the typical reason why commercial speech can be subject to greater governmental regulation than noncommercial speech."[97]

Analysis of an intermediate standard of review was articulated four years later in *Central Hudson Gas & Electric Corp. v. Public Service Commission*,[98] where the Court announced a four-part test for determining the constitutionality of commercial speech regulations. Only the first element is relevant to the present inquiry—i.e., that the regulation of commercial speech which is "misleading" merits no heightened scrutiny.[99]

The proposition that misleading statements or "false advertising" should be regulable is grounded upon its harmfulness.[100] The assumption is that false or misleading utterances have a systematic tendency to generate negligible social value—there being no good in promoting false beliefs—and a substantial negative externality arising from the disappointment or injury caused to consumers. Joined with the harm-deterrence argument,[101] the received view asserts, when a category of speech satisfies conditions (i) and (ii), its regulation is justified.[102] Therefore, regulating commercial speech with the objective of deterring false advertising enjoys the support of a harm-based justification.

---

[96] 507 U.S. 410 (1993).

[97] *Id.* at 426.

[98] 447 U.S. 557 (1980).

[99] *Id.* at 564.

[100] *See, e.g.*, Bolger v. Youngs Drug Products Corp., 463 U.S. 60, 81 (1983) ("The interest in protecting consumers from commercial harm justifies a requirement that advertising be truthful . . . ."); Procter & Gamble Co. v. Amway Corp., 242 F.3d 539, 549 (5th Cir. 2001) ("[F]alse or misleading commercial speech should receive no protection, because commercial speech merely gives information to consumers about a producer's goods, and any false information either has no value or is harmful."); Bates v. State Bar of Arizona, 433 U.S. 350, 379 (1977) (justifying regulation of false advertising by attorneys).

[101] *See, e.g.*, Bates, 433 U.S. at 380 (acknowledging the balancing of "possible harm to society from allowing unprotected speech to go unpunished"); Edenfield v. Fane, 507 U.S. 761, 762 (1993) ("A governmental body seeking to sustain a restriction on commercial speech must demonstrate that the harms it recites are real and that its restriction will in fact alleviate them to a material degree.").

[102] Central Hudson, 477 U.S. at 563 ("The First Amendment's concern for commercial speech is based on the informational function of advertising. Consequently, there can be no constitutional objection to the suppression of commercial messages that do not accurately inform the public about lawful activity. The government may ban forms of communication more likely to deceive the public than to inform it.") (*citations omitted*).

4.1.7 Obscenity

The last speech exception, for which I furnish examples of harm-based justification, is obscenity. Here, the appeal to the harmfulness of the proscribed expression is relatively undisguised.

Obscene speech was of course commonly regulated in English law. Whether this practice was intended to carry over after ratification of the Bill of Rights was an open question. However, as I discuss above, the United States Congress was in the early years of the Republic generally disinclined to intrude upon speech—with the exception of the short-lived Alien and Sedition Acts—and there were no federal obscenity laws until the mid-nineteenth century.[103] Wholesale regulation of obscenity was not aggressively pursued until passage of the Comstock Act in 1873, which banned the mailing of "[e]very obscene, lewd, or lascivious, and every filthy book, pamphlet, picture, paper, letter, writing, print, or other publication of an indecent character," as well as any information or products which could be used for the purpose of contraception.

To be sure, the legislation's leading proponent, Anthony Comstock, offered only vague consequentialist justifications for regulating speech.[104] Victorian social norms, a prying preoccupation with the sexual activities of the others, and assorted moral and religious fixations seem to have constituted Comstock's primary motivations. His moral crusading might be regarded quaint today, if not for a significant contingent of religious ideologues committed to carrying his torch into the twenty-first century. Unfortunately, Comstock's prurient grandstanding should not be dismissed as merely comic, for moral crusaders of his ilk still exert some—if thankfully dwindling—influence in the political arena.

Like his modern-day successors, the nearest Comstock came to a consequentialist justification for the wide-ranging regulation of "morally corrupting" speech was to gesture vaguely at the corruption of children and dissolution of the family. It is difficult to address these amorphous speculations directly, for neither Comstock nor his successors have attempted anything resembling a rigorous argument for them. He merely insists—without evidence—upon dubious psychological conjectures.[105]

---

[103] *See*, *e.g*, An Act to Provide Revenue from Imports, ch. 270, §28, 5 Stat. 548, 566–67 (1842) (prohibiting the importation of obscene material from abroad); Post Office Act, ch. 89, §16, 13 Stat. 504, 507 (1865) (prohibiting the mailing of obscene materials to Union troops); Post Office Act, ch. 246, §13, 15 Stat. 196 (1868) (prohibiting the mailing of information pertaining to lotteries).

[104] For an excellently informative historical review on the regulation of obscenity, *see* Margaret A. Blanchard, *The American Urge to Censor: Freedom of Expression Versus the Desire to Sanitize Society—From Anthony Comstock to 2 Live Crew*, 33 Wm. & Mary L. Rev. 741 (1992).

[105] *See* Anthony Comstock, Traps for the Young (4th ed., 1883). This work is a singularly

I shall nevertheless attempt a rational reconstruction of Comstock's argument: it seems he takes as a premise that human beings are universally happier when partaking in lifelong monogamous relationships, spawning offspring, and practicing obeisance to religious norms. He also evidently believed that ordinary persons are liable to be tempted—easily tempted—to a less fulfilling existence mired in sin and debauchery. The law ought therefore intervene, to set children and wayward adults on a better path, obliterating any information which might entice the weak-willed to act against their self-interest.

Though I have attempted to render Comstock's claims as charitably as I think possible, it is difficult to take any step of the argument seriously. It is plainly a harm-based justification, though for reasons I will enumerate below, I am skeptical whether it was (or is) ever intended sincerely. Although my aim in this article is to rebut harm-based justifications for the regulation of speech, and Comstock's argument is a harm-based one, I should not want to lump Comstock's argument with the rest. It is especially bad, and warrants separate treatment.

First, there is no evidence that human beings are generally (much less universally) happier in lifelong monogamous relationships. The archaeological record strongly suggests that early humans were polyamorous.[106] Lifelong monogamous relations

---

fascinating historical artifact, which I would commend anyone interested in obscenity law to skim. It is a *tour de force* of moral righteousness and paranoia: Comstock wastes no time, writing from the very first chapter, "Evil thoughts, like bees, go in swarms. . . . There is something wonderfully strange in the rapidity with which youthful minds take up lewd thoughts and suggestions." *Id.* at 7.

In a later chapter discussing the "free-love trap" (essentially, all sexual relationships except marriage), Comstock analogizes pre-marital and extra-marital sex to a "stone trap" he devised as a child to crush squirrels and rabbits. The analogy may perhaps resonate with priggish psychopaths—I must confess the significance of the grotesque allusion escapes me. Of "free love," he preaches:

> It takes the word 'love,' that sweetens so much of earth, and shines so brightly in heaven, and making that its watchword, distorts and prostitutes its meaning, until it is the mantle for all kinds of license and uncleanness. It should be spelled l-u-s-t, to be rightly understood, as it is interpreted by so-called liberals. . . . As advocated by a few indecent creatures calling themselves reformers—men and women foul of speech, shameless in their lives, and corrupting in their influences—we must go to a sewer that has been closed, where the accumulations of filth have for years collected, to find a striking resemblance to its true character. I know of nothing more offensive to decency, or more revolting to good morals, than the class of publications issuing from this source.

Id. at 158. Though Anthony Comstock's personage has faded somewhat from public notoriety, I hazard to suppose historians shall one day yet accord him his rightful place alongside the great moral purifiers—Savonarola, Torquemada, and Robespierre.

[106] *See generally* Christopher Ryan & Cacilda Jethá, *Sex at Dawn: How We Mate, Why We Stray,*

are an artifact of culture—not an inherent biological tendency. The preference for monogamy—to the extent that it exists—is not *innate* in the human species.

Of course, biology alone is not dispositive. Family values advocates will eagerly point to studies measuring the relative "happiness" of married versus unmarried persons, which they claim as evidence that traditional "family oriented" lifestyles tend to be more fulfilling. However, these surveys fail to disambiguate the influence of social norms and cultural expectations. Activists read the research as establishing the proposition that people tend to be happier in marriages. However, the data could just as well be read to reveal the perniciousness of social norms and expectations: that *nonconformity* with conventional social practices results in decreased happiness.[107] In other words, the data could be read as indicating not that marriage makes people happier, but rather that intolerant pro-marriage cultures cause unmarried people to feel less happy. Research on the discriminatory treatment of non-married people is still at a nascent stage, yet the evidence of insidious and pervasive mistreatment is compelling.[108]

The case for procreation as a universal private good is weaker still. *Even* in the face of pervasive social pressure to bear and raise children, there exists strong empirical evidence that individuals with children tend to be less happy than those without. More probing research on the effects of children on the welfare of parents is suspiciously difficult to find. Yet given the opportunity costs associated with having children, it is plausible to suppose that parents also suffer—in addition to psychological detriment—significant financial disadvantages from having children. Comstock's premise—that people tend to be "better off" having children—is wholly contradicted by the evidence. There may of course be *other* public policy arguments for encouraging procreation. I am doubtful however that anyone could reasonably argue that procreation is better *for parents*.

Finally, even if we were to grant *arguendo* that people tended to be more satisfied in lifelong monogamous relationships producing children, there is no evidence for the causal inference that individuals are as easily led to deviate from welfare-maximizing choices as Comstock assumes. The Comstock argument presupposes of humankind a degree of such abject stupidity, that even the hint of an "unwholesome"

---

and What It Means for Modern Relationships (Harper 2011).

[107] Moreover, no studies of which I am aware control for omitted variable bias.

[108] *See, e.g.* Bella M. DePaulo & Wendy L. Morris, *The Unrecognized Stereotyping and Discrimination Against Singles*, 15 Current Directions in Psych. Sci. 251 (2006); Wendy L. Morris et al., *No Shelter for Singles: The Perceived Legitimacy of Marital Status Discrimination*, 10 Group Processes & Intergroup Relations 457 (2007); Wendy L. Morris et al., *Singlism—Another Problem That Has No Name: Prejudice, Stereotypes and Discrimination Against Singles*, in The psychology of modern prejudice 165 (M.A. Morrison & T.G. Morrison, eds.) (2008).

thought would entice countless innocents to self-destruction. Even conceding that human decision-making is frequently and systematically susceptible to cognitive bias,[109] there is no principled reason to suppose "unwholesome" choices to be the product of cognitive bias rather than straightforward preference satisfaction. If lifelong monogamous relationships producing children were so clearly utility-maximizing, then the law would not need to enforce it. People would gladly pursue it of their own accord.

Curiously, Comstock denied that the regulation of obscene speech was speech regulation at all. In a glib move—frequently rehearsed by his successors—Comstock claimed that under his namesake legislation citizens would be free to express any thought or idea they pleased, so long as the expression were a legally permissible one. Thus, it was no limitation of the speech right to censor obscene expressions. It amounts to little more than saying that people are free to express what they please, except when they are not. Obviously, by this reasoning, the freedom of expression would be operative under *any* set of laws, no matter how restrictive. It is astounding how frequently this patently circular rhetoric is repeated.

Regardless of this, I expect that the modern day proponents of Comstock's position will find my counterarguments unsatisfying, because the harm-based arguments they propose were never what truly motivated their position. The consequentialist argument is for them an afterthought. The true ground for Comstockery seems not to be the regulation of harm, but rather the enforcement of morals. The reason why I expect my counterarguments against their harm-based justification would be unlikely to persuade them is because the harm-based justification was never the *real* justification for their position.

Nevertheless, it is revealing that Comstock and his successors felt the need to construct a consequentialist facade. They seem to intuit that nakedly advocating for the enforcement of morals would fail to persuade legislators, judges, and the public. There is implicit in their pseudo-harm-based justification a recognition that the law would require an identification of harms.

The earliest court opinions in cases challenging the Comstock Act ignored its First Amendment implications.[110] Disputes about its enforcement hinged upon procedural issues and precisification of what kinds of materials constituted obscene expressions.[111] The test which emerged was imported from the British case, *Regina*

[109] *See* Daniel Pi, *Meta-Rational Choice*, draft available at: https://ssrn.com/abstract=3226242.

[110] *See*, *e.g.*, United States v. Bennett, 24 F.Cas. 1093 (1879); Rosen v. United States, 161 U.S. 29 (1896).

[111] Skepticism about the semantic project was immediate and widespread, with Judge J.C. Ruppenthal neatly summing up:

> From the foregoing it may be seen that no general principle runs through the statutes

v. Hicklin.[112] For an expression to be obscene, it must "deprave and corrupt those whose minds are open to such immoral influences, and into whose hands a publication of this sort may fall."[113] Deferring to the legislature, the courts refrained from undertaking any constitutional review of the *justification* for the obscenity exception. To the extent the courts averred any justification at all, it was to acknowledge the propensity of obscene material to "corrupt the morals" of the citizenry (it is baffling that the courts seem unquestioningly to have regarded this the business of government to do).[114]

Such a condition could not stand, of course, and the courts have subsequently devised harm-based justifications—albeit ex post facto—to support the obscenity exception. The first test was articulated in *Roth v. United States*.[115] Writing for the majority, Justice Brennan invalidated the *Hicklin* test writing:

> The Hicklin test, judging obscenity by the effect of isolated passages upon the most susceptible persons, might well encompass material legitimately treating with sex, and so it must be rejected as unconstitutionally restrictive of the freedoms of speech and press.[116]

Brennan articulated the new standard: "whether to the average person, applying contemporary community standards, the dominant theme of the material taken as a whole appeals to prurient interest."[117]

In justifying the new standard, Brennan seems to reject the need for harm-based justification, writing:

> It is insisted that the constitutional guaranties are violated because convictions may be had without proof either that obscene material will perceptibly create a clear and present danger of antisocial conduct, or will

---

> of all the states, etc. As with laws everywhere that impinge upon sex matters in any way, there is more of tabu and superstition in the choice and chance, the selection and caprice, the inclusions and exclusions of these several enactments than any clear, broad, well-defined principle or purpose underlying them. Without such principle, well-defined and generally accepted, the various laws must remain largely haphazard and capricious.

Criminal Statutes on Birth Control, 10 J. Crim. L. & Criminology 48, 50 (1919).

[112] L. R. 3 Q. B. 360 (1868).

[113] Bennett, 24 F.Cas. at 1104 (quoting *Hicklin*).

[114] Judge Blatchford uses the phrase "corruption of morals" fourteen times in Bennett. Id.

[115] 354 U.S. 476 (1957).

[116] *Id.* at 489.

[117] Id.

> probably induce its recipients to such conduct. But, in light of our holding that obscenity is not protected speech, the complete answer to this argument is in the holding of this Court in *Beauharnais v. People of State of Illinois*, [343 U.S. 250, 266 (1952).]
>
> Libelous utterances not being within the area of constitutionally protected speech, it is unnecessary, either for us or for the State courts, to consider the issues behind the phrase "clear and present danger." Certainly no one would contend that obscene speech, for example, may be punished only upon a showing of such circumstances.[118]

Yet while denying the requirement of a harm-based justification, he contradictorily proceeds to reason in terms of harm-based justification. Brennan first asserts that exceptions to the general speech right must satisfy condition (i), writing:

> All ideas having even the slightest redeeming social importance—unorthodox ideas, controversial ideas, even ideas hateful to the prevailing climate of opinion—have the full protection of the guaranties, unless excludable because they encroach upon the limited area of more important interests. But implicit in the history of the First Amendment is the rejection of obscenity as utterly without redeeming social importance. This rejection for that reason is mirrored in the universal judgment that obscenity should be restrained . . . .[119]

And with respect to the particular controversy of the case, he emphasizes the likely failure of condition (i) to obtain, writing:

> Sex, a great and mysterious motive force in human life, has indisputably been a subject of absorbing interest to mankind through the ages; it is one of the vital problems of human interest and public concern[120]

Thus, despite asserting the absence of a harm-based justification, Brennan justifies invalidating the *Hicklin* test on the basis of overbreadth for likely failure of conditions (i) and (ii) to obtain in a substantial number of instances.

The cases which followed reveal the courts struggling to make sense of the *Roth* standard. The confusion was largely resolved in *Miller v. California*,[121] adopting an unequivocally harm-based justification for the obscenity exception, and setting

[118] *Id.* at 486–487.
[119] *Id.* at 484–485.
[120] *Id.* at 487.
[121] 413 U.S. 15 (1973).

the standard which remains the basis for determining whether expressions fall under the obscenity exception. Justice Burger, writing for the majority, established the three-part test:

> The basic guidelines for the trier of fact must be: (a) whether the average person, applying contemporary community standards would find that the work, taken as a whole, appeals to the prurient interest; (b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.[122]

The third factor of the *Miller* test ensures that condition (i) is satisfied, and the first two factors ensure that condition (ii) is satisfied. If a work "lacks serious literary, artistic, political, or scientific value," then it presumptively has negligible social value. And if a work is "patently offensive," adjudged by "contemporary community standards," then it presumptively generates a substantial negative externality. Thus, works which satisfy the three factors are potentially "harmful speech" and may be subject to government regulation.

4.2 Refutation of the Harm-Based Argument

The foregoing section establishes that throughout history, the *harmfulness* of a category of speech was indeed adduced as a justification for its regulation. In this section, I present my counterargument to harm-based justifications. Recall the harm-deterrence argument is implied in support of premise (2):

2.1. If harmful speech is punished, then the supply of harmful speech decreases.

2.2. If the supply of harmful speech decreases, then social welfare is improved.

2. If harmful speech is punished, then social welfare is improved. (*from* 2.1, 2.2)

My objective is to show that sub-premises (2.1) and (2.2) are false.

4.2.1 Endogenizing the *Harm* of "Harmful Speech"

Consider the inferential move in (2.2): that decreasing the supply of harmful speech will tend to effect an improvement in social welfare (or equivalently, a reduction in

[122] *Id.* at 24 (*citations omitted*).

total harm). The assumption driving the received view is that the average magnitude of harm caused by any particular harmful expression is *constant*, such that total harm varies with the quantity of harmful expressions. In other words, if one harmful expression generates social cost c, and the supply of a harmful type of expression is σ, then the total social cost is simply the harm multiplied by quantity, $c \times \sigma$.

I contend that this formulation is too simplistic. In very many circumstances, the harmfulness of an expression will decrease as supply increases. In other words, the average per-unit harm h will be a *function* of the quantity of expressions σ, such that marginal harm decreases as supply increases.[123] Thus, in case the marginal harm varies with supply, the social cost will *not* be $c \times \sigma$, but rather $h(\sigma) \times \sigma$.

The question now arises whether this is a better characterization of harmful speech in fact. For the sake of organizational clarity, I postpone full discussion of this issue to §3 below. However, I should remark in passing the facial plausibility of this premise. Consider obscenity as an exemplar. The intuition is that increasing exposure to "obscene" speech will tend to render it commonplace and less scandalous. The phenomenon is easily observed in the divergence between American and European attitudes toward nudity.[124] The relationship seems to be that the more common a putatively "obscene" expression is within a community, the weaker the negative reaction (if any). The intuition is easy to accept: that psychic phenomena like offense, credulity, outrage, or alarm obey the law of diminishing marginal returns. It is a lesson we all recognize from Aesop's *Boy Who Cried Wolf*.

Now, if within a given interval $\sigma \in I$, σ increases at a slower rate than $h(\sigma)$ decreases,[125] then it follows trivially that decreasing the supply of harmful expressions in that interval will have the effect of *increasing* total harm.

Consider the following numerical example. Suppose that $h : \sigma \mapsto 400\sigma^{-2}$. This satisfies the relevant conditions.[126] The following table, representing the values re-

[123] I.e., $\frac{\partial h}{\partial \sigma} < 0$.

[124] *See, e.g.*, Terrence Witkowskia & Joachim Kellner, *Convergent, Contrasting, and Country-Specific Attitudes toward Television Advertising in Germany and the United States*, 42 J. Busn. Res. 167, 171, 172 (1998) (finding Americans were much less likely than Germans to regard television advertisements as depicting a "wholesome" world, despite occasional frontal nudity in German advertisements); Colin R. Harbke & Dana F. Lindemann, *Acceptance of Female Public Toplessness: Structural, Contextual, and Individual Predictors of Support* 27 Canadian J. Hum. Sexuality, 92, 97 (2018) (finding 18% of Americans absolutely opposed to female public toplessness, as compared with 0% of Europeans).

[125] I.e., if $\frac{\partial h}{\partial \sigma}\sigma < -h(\sigma)$ obtains. To see that this condition results in a decrease in social welfare, observe that social welfare is decreasing iff $\frac{\partial}{\partial \sigma}\left(h(\sigma)\sigma\right) < 0$. Therefore, $\frac{\partial h}{\partial \sigma}\sigma + h(\sigma) < 0$, which is equivalent to $\frac{\partial h}{\partial \sigma}\sigma < -h(\sigma)$.

[126] I.e., $\frac{\partial h}{\partial \sigma} = -800\sigma^{-3} < 0$ for all $\sigma \in [0, \infty)$; and $\frac{\partial h}{\partial \sigma}\sigma = -800\sigma^{-2} < -400\sigma^{-2} = -h(\sigma)$ for all $\sigma \in [0, \infty)$.

turned by the function, illustrates the point:

Supply σ	Average Per-Unit Harm $h(\sigma)$	Total Harm $h(\sigma)\sigma$
5	16	80
10	4	40
20	1	20

Observe that although the supply of harmful speech σ increases, the average per-unit harm $h(\sigma)$ decreases at a faster rate, and therefore total social harm $h(\sigma) \times \sigma$ will decrease as the supply of harmful speech increases.

Of course, the numerical example is not proposed to model any specific real-world class of expressive activity. The function $h : \sigma \mapsto 400\sigma^{-2}$ was chosen to illustrate a mathematical point. In some real-world contexts, the harm function may be constant or increasing, in which case sanctions on harmful speech would, *ceteris paribus*, effect an improvement in social welfare. But the point is that situations comparable to the numerical example are *conceivable*. And if such conditions are *conceivable*, then premise (2.2) is false, and the harm-deterrence argument is unsound.

As a practical corollary, *even when* $\frac{\partial h}{\partial \sigma}\sigma \geq -h(\sigma)$, so long as $\frac{\partial h}{\partial \sigma} < 0$, the marginal effectiveness of legal sanctions will be mitigated to some extent. In such cases, even though decreasing the supply of harmful speech *would* decrease the social cost arising from that speech, it may still be inefficient to sanction harmful speech when we factor in the costliness of detection, sanctions, and litigation costs.

To restate the foregoing discussion less formally, the intuition is this: the average *potency* of some types of harmful speech diminishes as instances of such speech become more commonplace. Inversely, the average potency of such expressions increases as instances become scarcer. It is possible therefore that the average potency of expressions increases faster than the quantity of expressions decreases, in which case total harm is greater when there are fewer instances of harmful speech. Granting "harmful expressions" are net harmful, the tradeoff is that fewer expressions entail greater per-unit harm, and more expressions entail less per-unit harm. The question is what the total effect is. It is conceivable that a decrease in the quantity of harmful speech actually *increases* the total harm arising from that category of speech by increasing per-unit harm. As shorthand, let us refer to this phenomenon—where decreasing supply increases potency—as the "potency effect."

4.2.2 Endogenizing the *Benefit* of Harmful Speech

I next counter premise (2.1): the proposition that if harmful speech is punished, then the supply of harmful speech decreases. This premise may be further analyzed as

relying upon the principle that the imposition of legal sanctions on certain kinds of speech will tend to decrease the expected value of expression for prospective speakers of the relevant class. Assume that the prospective speaker of some harmful speech would derive some value from expressing it.

We may analyze the value the speaker derives from harmful speech into two components. First, the speech could generate value simply because the *expression itself* is gratifying to the speaker. Possibly it serves some innocent non-harm-directed purpose. Or possibly it is "therapeutic." Second—and more interestingly—the speech could generate value because the *harmful effect* it causes is desirable to the speaker. In other words, the "harm" is not merely incidental, but the very object of the speaker's intention. For convenience, let us call the first component "intrinsic value," and the second component "extrinsic value." I am principally concerned with the latter component.

Extrinsic value is a function of the magnitude of the externality. The greater its impact, the more value the speaker derives from engaging in it. This characterization is eminently plausible in speech contexts where the purpose of an expression is to shock, cause hurt, insult, or scandalize. In the category under consideration, extrinsic value V_E increases as the average per-unit harm h increases.[127]

The argument for the received view ignores the extrinsic value of harmful speech. Accounting for extrinsic value illuminates a second countervailing effect, which at a minimum weakens the argument for legal sanctions (and entirely nullifies it in the extremum case). For the purpose of illustration, suppose that the imposition of legal sanctions has the desired consequence of reducing the supply of some type of harmful speech. If the supply decreases, then assuming some level of potency effect,[128] the reduction in supply will effect an increase in the average per-unit harm of a given type of expression. This increases the extrinsic value of speech to prospective speakers, and thereby *increases incentives to produce harmful speech*, counteracting the initial reduction in supply.

Thus, in cases where potency effect and extrinsic valuation are present,[129] speech regulation operates like a spring. When the "force" of legal sanctions is applied in one direction, it is met with a proportional "force" in the opposite direction.[130] In the extremum case,[131] increasing sanctions will have no effect whatever on the supply of

[127] I.e., $\frac{\partial V_E}{\partial h} > 0$.

[128] I.e., $\frac{\partial h}{\partial \sigma} < 0$.

[129] I.e., when $\frac{\partial h}{\partial \sigma} < 0$ and $\frac{\partial V_E}{\partial h} > 0$.

[130] Although, *unlike* a physical spring, there is no particular reason to suppose the relation must be linear. Yet I am informed by my physics-degree wielding research assistant that not all springs respond linearly to force in any case.

[131] I.e., where $\frac{\partial V_E}{\partial h}\frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial S} > 1$, with S denoting the sanction level and $\frac{\partial \sigma}{\partial S} \leq 0$. In other words, if the

harmful expressions.[132] This defeats the claim that the imposition of legal sanctions on harmful speech necessarily decreases the incentives of prospective speakers to engage in it.

This general principle may also be observed in non-speech contexts. For example, in the worldwide market for elephant ivory, it has been argued that the imposition of legal sanctions on the international ivory trade, a policy adopted by 183 countries,[133] has had the perverse effect of decreasing the supply of ivory, thereby increasing its market value, and thereby increasing incentives for poachers to hunt elephants.[134] The underlying principle is essentially the same here.

Again, we have a practical corollary. In cases where the marginal effect of sanctions on the supply of harmful speech is *de minimis*, the costliness of detection, sanctions, and litigation can tip the balance, such that the net effect could be a *reduction* in social welfare when all factors are considered.

Restating the foregoing discussion less formally, the intuition is this: the speakers of some classes of harmful speech *intend* their expressions to cause harm. The imposition of sanctions on that class of speech reduces the payoffs that speakers receive from expressing it, reducing the supply of harmful speech. However, if the reduction in the supply of such speech increases hearers' sensitivity to that speech, they will be more susceptible to being harmed by it. This increases the benefit that speakers receive from expressing it, thereby increasing the supply of harmful speech. In the limiting case, the countervailing effect entirely cancels out the reduction in the supply of harmful speech. As shorthand, let us refer to this phenomenon as the "inverse supply effect."

4.2.3 Filtering

For the third countervailing effect, I return to premise (2.2): that if the supply of harmful speech decreases, then social welfare is improved. Assuming the first two effects are present for some class of expressive activity, there arises the possibility of

effect of increasing sanctions decreases supply, and decreasing supply increases marginal harm, and increasing marginal harm increases marginal extrinsic value, then the net effect on the speaker's incentive could, in the limiting case, be null.

[132] I have in mind a formulation of the speaker's utility function as being $U = V_I + V_E - S$, where V_I denotes the intrinsic value of speech. This formulation ignores imperfect enforcement—without loss of generality.

[133] The relevant treaty is the Convention on International Trade in Endangered Species of Wild Fauna and Flora, Mar. 3, 1973, 27 U.S.T. 1087, 993 U.N.T.S. 243.

[134] For example, see Daniel W.S. Challender & Douglas C. MacMillan. "Poaching is More than an Enforcement Problem." 7 Conservation Letters. 484–494 (2014).

a selection effect, which further undermines the premise.

If we assume that prospective speakers are heterogeneous in the magnitude of harm their speech causes, then the imposition of legal sanctions may not only affect the *size* of the population of speakers, but also the *composition* of that set. Let us partition the population of speakers into two subsets: high-harm speakers and low-harm speakers.[135] The distinction is that high-harm speakers are those who—whether by circumstance or skill—are likely to cause greater harm than low-harm speakers when engaging in a type of speech. The construction is open to the possibility that an individual may be a high-harm speaker in one situation, but a low-harm speaker in another. For example, a given person who is artless in lying (and thus a low-harm speaker with respect to libel) may have a talent for causing offense (and thus be a high-harm speaker with respect to obscenity).

Now, suppose the conditions for the potency effect and the inverse supply effect are satisfied.[136] It follows that low-harm speakers will be more susceptible to deterrence than high-harm speakers, because they derive less benefit from causing harm. This is definitional. A low-harm speaker is one whose harmful speech causes less harm than that of a high-harm speaker. Thus, the extrinsic benefit that the low-harm speaker derives must be less than the extrinsic benefit that the high-harm speaker derives, because the harm *caused* is less. Recall, extrinsic value is defined as increasing as the harm caused increases. Therefore, assuming the sanction level is held constant for all individuals, low-harm speakers will, *ceteris paribus*, be deterred more easily than high-harm speakers.

But if low-harm speakers elect to refrain from engaging in harmful speech, then they will have decreased the supply of total speakers, and by the potency effect, increased the marginal harm experienced by hearers. And by the inverse supply effect, this increases the incentives of speakers to engage in harmful speech. Specifically, it will tend to be *high-harm* speakers, less sensitive to the deterrent effect of sanctions, who are incentivized to "fill the gap" left by the low-harm speakers who were deterred.

Thus, *even if* the benefit of regulating speech outweighed the potency effect and inverse supply effect, it may *still* be the case that speech regulation would be inefficient. Although the supply of harmful speech may be reduced, nevertheless the altered composition of the population of speakers might result in an *increase* in total harm. The principle is intuitive and may be better communicated with the aid of

[135] We might formalize this by defining two distinct harm functions, such that the high-harm speaker's expressions generate h^H harm, and the low-harm speaker's expressions generate h^L harm, such that $h^H(\sigma) > h^L(\sigma)$ for all σ.

[136] I.e., $\frac{\partial h}{\partial \sigma} < 0$ and $\frac{\partial V_E}{\partial h} > 0$.

a numerical example. The following table illustrates the point with hypothetical values. Assume that the average per-unit harm of low-harm speech is 5, and the average per-unit harm of high-harm speech is 10.

	Total Supply of Harmful Speech	Low-Harm Speech	High-Harm Speech	Total Harm
Unregulated	10	7	3	65
Regulated	9	4	5	70

Note that these numbers were chosen not to represent a specific real-world effect, but rather to develop an intuition for the general principle. As shorthand, let us refer to this phenomenon—where the imposition of sanctions screens out low-harm individuals, increasing incentives for high-harm individuals to fill their place—as "filtering." Again, in the interest of organizational clarity, I discuss the plausibility of filtering in real-world circumstances in §3 below.

To be sure, filtering is the most tenuous of the three countervailing effects. It is tenuous, firstly, because it *assumes* the presence of the other two effects. Whereas the potency effect may arise even when the conditions for the inverse supply effect and filtering are not satisfied, the inverse supply effect depends upon the presence of the potency effect. And filtering depends upon the presence of *both* the potency effect and the inverse supply effect. Secondly, filtering requires additional assumptions, which are disputable. It assumes that high-harm speech and low-harm speech of a given type are substitutes, differing only in magnitude. It also assumes that high-harm individuals are, apart from the definitional distinction, otherwise identical to low-harm individuals. However, it may be that low-harm individuals inflict low magnitude harm, because they differ in motivation. For example, it may be that high-harm individuals inflict high magnitude harm because their speech is principally *motivated* by a desire to cause harm, whereas low-harm individuals inflict harm only incidentally. If the reason why low-harm individuals inflict low magnitude harm is because they are motivated by *other* objectives, then the comparatively lower extrinsic value they derive (relative to high-harm speakers) may be offset by higher intrinsic value (i.e., their principal motive for engaging in such speech—whatever that might be). If this were the case, then my claim that low-harm individuals are more susceptible to deterrence would be mistaken. Nevertheless, I think filtering *could* arise in many speech contexts, and recognition of the possibility is sufficient to establish the falsity of premise (2.2).

And once again, there is the corollary that *even if* regulation were welfare-improving, despite the presence of all three countervailing effects, extending the model to account for the costliness of detection, sanctions and litigation can still render speech restrictive policies inefficient in practice.

4.2.4 Counter-Argument in General Form

In the preceding subsections, I have identified three countervailing effects, which undermine the argument for restrictions on harmful speech. The point is that merely identifying a class of expressive activity as satisfying conditions (i) and (ii) does not suffice to justify its regulation. Indeed, I identify conditions where regulation of harmful speech can even have the effect of *increasing* total harm.

To summarize: the *potency effect* identifies a tradeoff between the quantity of harmful speech and the average per-unit harm of that speech; the *inverse supply effect* identifies a friction which can undermine the effectiveness of sanctions on expressive activities, arising from increasing incentives to produce harmful speech when quantity decreases; and *filtering* identifies screening that could arise when speakers are heterogeneous in the magnitude of harm their expressive activity causes. Where speakers are analogized to producers and hearers are analogized to consumers, it may be helpful to think of the potency effect as a "demand-side" effect, and inverse supply and filtering as "supply-side" effects.

All three effects undermine steps in the argument for the received position. The conventional view is that imposing (or increasing) sanctions on harmful speech deters production of that speech, thereby effecting an improvement in social welfare. I demonstrate that the conclusion does not necessarily follow from these premises. Because premises (2.1) and (2.2) of the harm-deterrence argument can fail, it follows that premise (2) is not necessarily true, and therefore that the harm-based justification requires more than satisfaction of conditions (i) and (ii) to be sound.

It is important to distinguish what I am *not* saying. The presence of these effects does not establish that sanctions *cannot be* welfare-improving (or are *always* welfare-reducing). To assert this would be claiming too much. Rather, what the possible presence of these effects does is it undermines an essential inferential component of the harm-based justification. It reveals two premises to be less sturdy than previously supposed. This raises the argumentative bar for the regulation of speech. It requires proponents of speech restrictions do more than merely point to the net harmfulness of a category of expression to justify regulating it.

It is also important to distinguish how my argument differs from other counterarguments opposing the regulation of harmful speech. Very many of the counterarguments opposing the regulation of harmful speech contest whether a putatively "harmful" expressive activity is *truly* harmful—when all factors are considered. These counterarguments seek to undermine claims that a category of expressive activity satisfy conditions (i) or (ii). Although I am broadly sympathetic with this line of attack, this is not the nature of my counterargument here. My counterargument accepts, *arguendo*, the undesirability of putative "harms," but contends that the im-

position of sanctions may be *ineffective* at reducing that harm. Indeed, my argument suggests that sanctions may even exacerbate the harm in some circumstances.

Building on my negative argument, I have also hinted at a stronger, affirmative claim: that the presence of costly detection, sanctions, and litigation may tip the scale. *Even if* a speech regulation would reduce the direct harm from speech—despite the three countervailing effects—their presence would nevertheless mitigate the positive effects of such regulation. At this point, the enforcement costs (i.e., detection, sanction, and litigation costs), could push marginal regulation into the negative, so that it represented a social loss.

Yet aside from the enforcement cost "kicker," I have another affirmative argument still. If the countervailing effects are sufficient to nudge the balance of social welfare close enough to the indifference point—where we are uncertain whether the total effect is positive or negative—then I contend the presumption ought to be *against* regulation, for if there is any meaning at all to the general speech right, it must at a minimum establish a *default* rule in cases of uncertainty. If it is not even that, then it is nothing at all. Ergo, when it is unclear whether a speech regulation would effect an improvement or reduction in social welfare, the presumption ought to be against government interference. This is not only what the law *ought* to be, but what the law *is*. The principle is implicit in nearly every judicial opinion: that in the absence of a harm-based justification, the government may not regulate speech.

Thus, if the three countervailing effects succeed only in pushing the balance of social welfare to within the penumbra of uncertainty, two considerations will militate against the regulation of speech. First, the costliness of detection, sanctions, and litigation. Second, the default presumption of the general speech right against regulation.

4.2.5 Counter-Counterarguments Anticipated

I anticipate two potential objections. First, the critic might object that the three effects I have described heretofore are not *unique* to speech restrictions. I imagine such objections hitting upon the presence of the three countervailing effects in other realms of conduct, where a restrictive law is nevertheless deemed warranted. These objections would be in the class of *ad absurdum* arguments, possessing the following form:

> Some non-speech activity X exhibits one or more of the three effects. When some of the three effects are present, regulation may be inefficient. Therefore, regulation of X may be inefficient. Therefore X should not be regulated absent further evidence.

> *But* it is absurd to think that regulating X could be inefficient. Therefore the inference—that regulation in the presence of the three countervailing effects might entail a reduction in social welfare—must be invalid.

The strength of this counter-counterargument depends entirely upon the strength of the premise that regulating X is "obviously" efficient. Presumably, the proponent of speech regulation would choose an X that most people would readily agree ought to be regulated.

For example, consider if X were vandalism. The counter-counterargument would observe the possible presence of the potency effect, the inverse-supply effect, and filtering when regulating vandalism. Thus, if my argument were to be taken seriously, then it would follow that the law ought not prohibit vandalism. And yet, my hypothetical opponent would contend, it is *obvious* that the law should prohibit vandalism, and therefore the general principle of my argument fails.

I have three responses. First, assuming the wisdom of regulating X is obvious, I would question whether one really could not satisfactorily provide the additional reasons which justify the inference to premise (2). In the case of vandalism, I think it may be plausibly argued that decreasing the quantity of vandalism would not necessarily increase the average per-unit harm of the undeterred residual vandalism. In this case, it may simply be observed that the potency effect is *not* present, and since the inverse-supply effect and filtering depend upon the potency effect, we may conclude that none of the three countervailing effects obtain in fact.

However, if it were established empirically that the potency effect were present for some X, then my second response would be to inquire whether there exist kickers militating against regulation of X. Again, in the case of vandalism, there does not seem to be any general category of behavior, of which vandalism is a subset, which the law accords as a general right. Also, it seems that enforcement costs may be offset by the cost of repairing the vandalized property. Therefore, even if it were uncertain what effect vandalism regulations would have on social welfare, there would be no affirmative reason *not* to regulate it.

And finally, *if* it were established that the potency effect were present for some X, and *if* there were a general right to some broader class of activities which included X, then my response would be to embrace the "absurdum" conclusion. Supposing it were shown that vandalism regulations exhibited the three countervailing effects, and supposing there were some general right that vandalism regulations would incur upon, I would incline to accept the conclusion that vandalism ought not be prohibited. I am skeptical of course that either of the conditions could be met in the case of vandalism particularly, but if it could be shown that the three countervailing effects were truly present, and the balance of social costs were unclear, and there existed kickers which

tended to militate against regulation, I should think it not at all "obvious" that X should be regulated.

The second objection I anticipate is more general. Some readers may complain that my negative argument proceeds from armchair observations about a theoretical model unsupported by empirical data. The hostile critic will protest that I do not even attempt to determine the values of variables nor to discern concrete facts.

This grossly misses the point. As I demonstrate in §1, the conventional justification for speech regulation relies upon an economic inference—the harm-deterrence argument. The harm-deterrence argument depends upon premises which are theoretical economic claims. The counterargument which I provide is correspondingly a theoretical economic counterargument.

Complaining that my counterargument fails to delve into empirics mistakes the logical structure of my contention. To illustrate the point, suppose a person asserted the argument: all cats are quadrupeds; all dogs are quadrupeds; therefore all cats are dogs. This is clearly a textbook "illicit minor" fallacy. The proper counter here is to demonstrate the invalidity of the inference: that "all X are Y" and "all Z are Y" does not entail "all X are Z." The counter requires no empirical investigation into the nature of cats, dogs, quadrupeds, or any other possible X, Y, or Z.

Similarly, the received justification for speech regulation relies upon an economic inference. The presence of the three countervailing effects undermines that inference. No empirics are necessary to demonstrate the unsoundness of the inference. Indeed, empirical claims are entirely *irrelevant* to the point in dispute. To insist otherwise is to misunderstand the nature of scientific theory generally.

4.3 Applications

My objective in this article being to show that the harm-deterrence argument is unsound, it suffices to demonstrate the possible occurrence of the three countervailing effects to establish my thesis. However, some intuitive connection to real world controversies may be wanted. I expect both proponents and opponents of speech regulation will be interested not only in whether the three countervailing effects are possible, but whether they are plausible. In other words, whether there is a coherent intuitive story where the three countervailing effects are believably present in areas of controversy.

In this section, I consider several prominent examples, supplying reasons to believe that potency, inverse supply effect, and filtering might arise. I do not intend these examples to be exhaustive. My counterargument is deliberately framed in general terms, and I expect the three countervailing effects might arise in any cir-

cumstance where harm-based justifications for regulation are proposed. Neither do I intend the examples to be conclusive. A rigorous empirical investigation into any of the particulars would require considerably more than I could hope to furnish within the scope of this article. My aim in this section is simply to sketch out the facial plausibility of the effects obtaining in a variety of contexts.

4.3.1 Affronting Speech

Let us start by considering a subset of harmful speech, the supposed harm of which is psychic injury. Let us refer to this subset as "affronting speech." Affronting speech represents the simplest and most straightforward context to which my argument applies.

There are two relevant player categories in the context of affronting speech: injurers and victims. The conventional justification for the regulation of affronting speech is that injurers impose negative externalities upon victims, reducing social welfare. Regulation is meant to deter injurers from expressing harmful speech. To reiterate: I take the characterization—that the relevant *harm* is the psychological effect it has upon victims—as defining the subset of speech which is affronting.

Affronting speech is analyzable into familiar doctrinal types. I include in the category of "affronting speech": offense, fighting words, hate speech, obscenity, indecency, and profanity.

Offense

In the context of offensive speech, the relevant "affront" is the displeasure experienced by some portion of the public when a cherished symbol or idea is debased. Typical examples include the desecration of flags or religious symbols.[137]

[137] An especially broad curtailing of speech rights on the basis of offense may be observed in the jurisprudence of Poland. Curiously, the Polish Supreme Court has ruled that the freedom of religion includes the right of religious people "not to be offended."

> "The subject of the protection to which Article 196 of the Criminal Code pertains are the religious feelings arising from the constitutional freedom of religion. It is accepted in the jurisprudence that the religious feelings are, due to their nature and the direct link with freedom of religion, subject to special protection. ... There is no doubt that the protection of religious feelings, and thus human emotion associated with the faith professed by the individual, is also linked to the protection of the inherent and inalienable dignity of the human person, which is the source of freedom and human and civil rights [citing Article 30 of the Polish Constitution] ... Article 196 of the CC serves as an expression of a specific position taken by the legislature in the potential conflict of freedom of expression ... The legislature decided the conflict in favor of

Here, the presence of all three countervailing effects may be observed. First, it is surely the case that marginal harm decreases as quantity increases. To take the most notable exemplar of "offensive speech," observe that public sentiment toward the proposal for a constitutional prohibition on flag-burning has, since *Texas v. Johnson* in 1989,[138] steadily waned.[139] Without intending to diminish the possible contribution of other causal factors,[140] I would nevertheless expect the magnitude of popular support for a "Flag-Burning Amendment" to be a reasonable proxy for the magnitude of psychic harm that the public experiences when witnessing immolation of the flag.

This is of course a thoroughly unsurprising fact. It cannot reasonably be doubted whether an expression which elicits hostility will tend to diminish in potency as individuals susceptible to offense grow inured to the stimulus through repeated exposure. Common sense and everyday experience confirm this phenomenon.

Second, the inverse-supply effect requires that the benefit to injurers increase as harm increases. This is plausible for offensive speech. Returning to our exemplar, it is surely the case that those who burn flags do so precisely *because* they anticipate the offense it will cause. Indeed, it is difficult to imagine any *other* reason to endure the expense, risk of injury, and noxious fumes which the act entails, except to agitate the passions of those spectators who witness the incineration of their beloved symbol.

> freedom of religion, assuming that it is unlawful to express the views that consist of insulting of religious object or place intended for public performance of religious ceremonies, which leads to offending religious feelings of others."
>
> Judgment of the Constitutional Tribunal of October 6, 2015 (Case no. SK 54/13).

[138] 491 U.S. 397 (1989) (holding the burning of the American flag to be constitutionally protected speech). *See also* United States v. Eichman, 496 U.S. 310 (1990).

[139] A Gallup poll taken in 1989 found 71% of respondents in support of a Constitutional Amendment prohibiting flag burning. By 1990, this number had declined to 68%. By 1995, those in favor had declined to 62% (the support level held in 1999 at 63%). By 2006, Gallup found only 56% in favor of such an Amendment. Carroll, Joseph, "Public Support for Constitutional Amendment on Flag Burning." Gallup News Service. June 29, 2006, *available at:* http://news.gallup.com/poll/23524/public-support-constitutional-amendment-flag-burning.aspx. The most recent numbers—the 2011 State of the First Amendment Report, *available at:* http://www.newseuminstitute.org/wp-content/uploads/2014/09/FAC_sofa_2011report.pdf—suggest that a majority of Americans now oppose a flag burning Amendment, with only 39% in favor. It is worth noting that the numbers available from Gallup and the State of the First Amendment Reports do not agree on the years where they overlap. Nevertheless *both* surveys demonstrate clear diminishing support for a Flag-Burning Amendment.

[140] For example, I imagine that the expressive effect of the law would be pronounced in these circumstances. The expressive factor—although not identical to the potency effect—would tend to reinforce my arguments regardless.

Unfortunately, there do not exist good statistics on the frequency of flag-burning. Yet even if there were, the other relevant variable—i.e., the magnitude of outrage—would remain difficult to quantify. It is telling however that the incidence of flag-burnings seems to increase precisely during those periods when proposals for flag-burning prohibitions are prominent in public discourse.[141] The economic explanation is obvious: proposing to ban the burning of flags is tantamount to a declaration that one is acutely susceptible to its intended effect. It signals to injurers a high "price" they can extract from offensive expression, which naturally elicits an increase in production.

Additionally, the inverse supply effect also has an interesting interplay with publicity. News outlets, motivated to attract a novelty-craving audience, have greater incentivizes to report on seemingly uncommon events. Thus, even if the imposition of sanctions were to reduce the supply of injurers, this may have the perverse effect of amplifying instances of a type of speech so as to affect more recipients. As the publisher Alfred Harmsworth once famously quipped, "When a dog bites a man, that is not news, because it happens so often. But if a man bites a dog, that is news." Correspondingly, when the "news" is something many people find offensive, the consequence of decreasing supply is increasing publicity, and with increasing publicity, increasing harm. The publicity effect can be regarded as a species of (or at least approximately equivalent to) the inverse supply effect.

Third, it seems plausible that filtering would also arise in the context of offense. Again, to take flag-burning as an example, we expect that the imposition of sanctions would naturally increase the expected cost of expression. Given that the inverse-supply effect obtains, injurers will tend to derive greater utility the more widely publicized their flag-burning, and less utility the less widely publicized. Thus, injurers will be deterred from burning the flag in circumstances where the harm is low (i.e., when there are fewer potential recipients/victims), reducing supply, and increasing the potency of the expression. The effect is that the grandiose act of flag-burning will be reserved for those occasions when it will have the most impact. The savvy protester will refrain from burning the flag when few passersby will notice, avoiding the disutility of sanction. He will hold his shock tactic in reserve until he has attracted the attention of photographers and journalists—to maximize the value he derives from incurring the sanction.

[141] For example, after President-elect Trump suggested that flag-burners should suffer criminal punishment (or a loss of citizenship), a small band of protestors gathered in front of Trump's New York Hotel to burn American flags. Stapleton, Shannon. "Trump flag-burning tweet leads activists to burn some flags in New York." *Available at:* https://www.reuters.com/article/us-usa-trump-flag/trump-flag-burning-tweet-leads-activists-to-burn-some-flags-in-new-york-idUSKBN13P06L.

It is thus eminently plausible that all three effects would arise in the regulation of flag-burning. And the intuitive story maps easily to offensive speech generally. Whether the vehicle for expression takes the form of burning flags, draft cards, or books, carrying offensive signage, or printing offensive slogans on tee-shirts, we should expect increasing the supply of such expressions will tend to reduce their impact. And it is a safe surmise to suppose the speakers derive utility from the disutility they cause in others. And increasing the cost of expression through sanction would tend to deter low-harm instances more than high-harm instances, as strategic speakers "reserve" their expression for those occasions when it is likely to have the most impact.

Obscenity, Indecency, and Profanity

Similarly, in the context of obscenity, indecency, and profanity, we may conceive the relevant affront to be the displeasure experienced by some portion of the public, when some "vulgar" expression is given voice. Typical examples include the portrayal of sexual activity, defecation, or urination; utterances of "dirty words"; depictions of nudity; devices and implements of sexual fetishists; and media tending to the subversion of social norms.[142]

Let us take the "shock art" movement as our first exemplar.[143] A cursory investigation of its history reveals robust examples of the potency effect. Among the principal developmental mechanisms of "shock art" seems to be simple one-upmanship. Each subsequent generation of shock artists seeks to unsettle conventions more radically than their predecessors had done. This may not be the *only* mechanism at work, however it is surely one of the main factors driving the genre.

The phenomenon follows from the principle: that scandals subside. It is the artist's analogue to the second law of thermodynamics. Entropy, it seems, can be discovered in many forms. Marcel DuChamp's once outrage-inducing *Fountain* (1917) sits today in the Museum of Modern Art in New York, evoking little reaction but the cool insouciance of perplexed tourists. The premiere of Stravinsky's *Rite of Spring* in 1913 induced a riot. It is now a staple of the orchestral repertoire; and its per-

[142]Curiously, this has, in many totalitarian regimes, included abstract artistic works with little or no representational content whatever. For example, [DISCUSS Entartete Kunst in NAZI GERMANY]. It is ironic that historically the most ferocious attacks on art have targeted those works which have possessed the least political content.

[143]Notable representatives of this "movement" include Andres Serrano, who photographed a crucifix submerged in his urine (*Piss Christ*, 1987), Chris Ofili, whose *Holy Virgin Mary* (1996) depicts the eponymous subject as a Black woman, with one exposed breast constructed from varnished elephant dung, against a collage of pornographic imagery, and Rick Gibson, who fashioned earrings from freeze-dried human fetuses (*Human Earrings*, 1987).

formance provokes little more than dumb indifference from geriatric concertgoers struggling with their hearing aids.[144] What is an artistic *affront* today is merely banal tomorrow. Its potency diminishes as audiences grow inured to its novelty.

Discounting the attestations of shock artists themselves (their propensity to use public statements to subvert expectations renders their putative self-reflections incredible), it certainly *seems* from observation of their behavior that they are deliberate in their attempts to elicit "shock" responses.

This is wholly consistent with economic principles. The shrewd artist, seeking novelty, attention, notoriety, and wealth, is wise to court controversy. Thus, reducing the quantity of shock art would tend to increase incentives for shock artists to create it. In other words, if the imposition of sanctions on obscene art reduced the supply of shock art, then by the potency effect it would increase audiences' sensitivity to shock, thereby increasing the incentives for shock artists to produce it, and by the inverse supply effect counteract the initial reduction in supply.

The conditions for filtering seem to be satisfied also. Low-harm artists, whose work might incidentally cause shock, or who lack talent for producing shocking art, would be more susceptible to deterrence, because they derive less benefit from their relatively less effective attempts to elicit outrage. If they exit the market, decreasing the supply of obscene art, then by the potency effect, audiences' sensitivity to obscenity will increase. And those artists whose comparative advantage lies in their talent for shocking will experience stronger incentives to produce obscene art.

All three effects are therefore plausibly present in the context of shock art. It follows that imposing sanctions on shock art could result in a net reduction in social welfare, undermining harm-based justifications for censorship of it. The critic may object that I am cherry-picking in my choice of exemplar—that I have deliberately chosen shock art, because it is an especially availing context in which to observe the three countervailing effects. This misses the point. Shock art is acutely relevant, because it is the sort of artistic expression most liable to be regulated. It is the frontline in the battle for free speech. I am happy to concede that the inverse supply effect is unlikely to arise if governments were to regulate, for example, landscape paintings or muzak. This is irrelevant. The implausibility of the countervailing effects arising from the regulation of landscapes and muzak follows directly from the implausibility of their causing harm in the first place.

A still more vivid illustration of the three countervailing effects may be observed

[144] I do not mean to imply that the *Rite of Spring* was conceived as "shock art." Arguably, no analogous movement ever arose in music—although George Antheil seems to have invested some effort in assuring that the premiere of *Ballet Mécanique* (1926) resulted in scandal, and of course John Cage also routinely courted controversy.

in the regulation of "pornographic" films. The earliest cinematic depiction of a remotely sexual nature is *The Kiss* (1896). One of the first commercially distributed films, *The Kiss* depicts a fully clothed man and woman nuzzling and exchanging brief pecks over the course of eighteen seconds (looped thrice). *The Kiss* was distributed by Thomas Edison in an effort to promote the kinetoscope—a nickel-operated, hand-cranked motion picture invention.[145]

The kiss depicted in *The Kiss* is not, by modern standards, a remarkably passionate osculation. Yet it was deemed reprehensible in the twilight of the nineteenth century, provoking one critic to write:

> The spectacle of the prolonged pasturing on each other's lips was hard to bear. When only life size it was pronouncedly beastly. Magnified to Gargantuan proportions and repeated three times over, it is absolutely disgusting.[146]

The Catholic Church denounced *The Kiss*, calling for censorship.[147] Outrage over the film had hardly subsided before Edison released a still more scandalous vignette, *Dolorita's Passion Dance* (1897), depicting a (once again fully clothed) woman engaging in an Iberian "passion dance." So intolerable were her bodily contortions it seems that the sage authorities of New Jersey were left with little choice but to raid the Atlantic City parlor, in which it was being shown, and to ban the film.[148]

The modern media consumer need only introspect, to seek out even a quantum of revulsion in his own mind upon viewing *The Kiss*, to recognize the potency effect at work. Indeed, I doubt modern viewers would recognize *The Kiss* as being erotic in any sense at all. Whatever harms contemporary moral figures may have experienced in the furor over *The Kiss* are wholly incomprehensible in the present day, and the complete dissipation of that harm is undoubtedly due to the overwhelming quantity of onscreen kissing in television and cinema.

Evidence of the inverse supply effect can also be observed in the early days of cinema. The notoriety of *The Kiss* inspired imitators, presumably seeking to exploit the titillating potential of the new medium.[149] Yet the erotic novelty of *The Kiss* and its copycats abated rapidly. While the demise of the kiss porn genre was likely due

[145] *See generally* Geltzer, Jeremy, *Dirty Words and Filthy Pictures: Film and the First Amendment* (2016) for a thoughtful historical survey of censorship in cinema.

[146] *Id.*

[147] *Id.*

[148] *Dolorita's Passion Dance* was thus the first censored work of cinema. It was also, at the time the mayor's order was issued, the "most viewed Kinetograph picture the parlor had ever hosted." Geltzer, 9, citing.

[149] *The Kiss in the Tunnel* (1899) and *The Kiss* (1900).

to the emergence of more extreme content,[150] the requisite principle is retrievable: that injurers derived a benefit positively correlated with the magnitude of offense caused.[151] As the sensitivity of audiences to onscreen kissing diminished, so too did incentives to exploit that effect.

It is true that onscreen kissing remains exceedingly common in films today, and the inverse supply effect implies that as audiences grow inured to the effect, incentives to depict it should decrease. Yet it is important to distinguish that the cinematic depiction of kissing was not exclusively or (except perhaps in the very earliest days of cinema history) even predominantly motivated by the pursuit of extrinsic value. The portrayal of kissing has substantial intrinsic value. It is a useful expository tool in the construction of stories—in depicting the relationships between characters, and revealing characters' emotional states visually. The speaker's extrinsic motivation—the benefit received from causing offense—in the depiction of kissing seems entirely absent in the modern filmmaker.

Likewise, filtering also seems to be present. Whereas mainstream filmmakers, for whom eroticism was but an incidental element of their art, sought to stay within (if only *just* within) the boundaries of what was permissible, the scofflaws whose very purpose was to produce lascivious content would exploit the regulatory effect of censorship, charging high prices for content which prevailing social norms would ensure possessed a high degree of potency.

Now, proponents of obscenity, indecency, and profanity regulation may counter that I have missed the point entirely. They will not dispute that the effect of pornographic imagery tends to dissipate with increased exposure. Indeed, this is the gravamen of their complaint. They will concede that liberal speech laws dilute the harm of *The Kiss*, yet they will insist that a society indifferent to "corrupting imagery" is somehow a *worse society*. The *harm*, they might argue, is not the affront, but rather the coarsening of standards which preempts affront. Their argument is that it would be as though a patient went to a doctor complaining of soreness in his arm, and the doctor "cured" the ailment by severing a nerve so that the patient lost all feeling that limb.

This counter does not touch my argument. Indeed, it concedes to it. The argument that a society *more* sensitive to harms is somehow a "better society" relies

[150] For example, Georges Melies' *Après Le Bal* (1897, Fr.), depicting a fully nude woman, and *A L'Ecu d'Or ou la Bonne Auberge* (1908, Fr.), depicting penetrative intercourse.

[151] The mechanism has one complication. The pornographer seems not to have been motivated to *cause* offense directly. There is an intermediate inference. Content which is sufficiently novel to stimulate erotic sentiments for a segment of the population will tend to be sufficiently novel to offend the indignation of the other. The effect will follow, even if the extrinsic valuation is not direct.

on premises which cannot easily be reconciled with a welfarist model. It is a moral argument masquerading as a consequentialist argument. And indeed I think its irreconcilability with a welfarist model is itself evidence that there is no substance whatever in the claim.

Nevertheless, I think a brief digression may be warranted on this flimsy contention. I have two responses. First, to the extent that proponents of the obscenity, indecency, and profanity exception would claim that censorship is something like a defense of a society's moral identity, they are taking up the losing side of a settled controversy. They revisit the arguments of Lord Patrick Devlin in his debate with H.L.A. Hart, in which Hart conclusively prevailed.[152] Second, it seems doubtful that any person today would honestly contend that society was somehow better off (even "morally" better off) when a grainy close-up shot of a poorly-aimed kiss constituted a scandal. Can any proponent of the obscenity, indecency, and profanity exception really believe that our present indifference to *The Kiss* is a detriment? The question may be extended to later targets of censorship. Would our society be better off without films like *Scarface* (1932) or *Monty Python's Life of Brian* (1979)? Would we be better off if Lenny Bruce or George Carlin were deterred from the expression of profane comedy? Would we be better off if the censors had prevailed in extinguishing James Joyce's *Ulysses* or Nabokov's *Lolita* from bookstores? I expect no reasonable person would answer these questions in the affirmative.

The modern proponent of the obscenity, indecency, and profanity exceptions may concede that these historical examples were unwarranted government intrusions, and that our society is no worse off for tolerating these former targets of speech regulation after all. But he may contend *those* obscenities were never *really* obscene in the first place. He will maintain that the obscenities, indecencies and profanities he perceives today are *different*. These things *really are* obscene, indecent, or profane. And *this time*—he will insist while marveling at the stupidity of his censorious predecessors in drawing the line so poorly—*this time* he has surely gotten it right.

There is no scientific answer which would satisfy the moral ideologue, though I am skeptical whether he deserves a response at all. Regardless, for present purposes it suffices to observe the plausibility of the three countervailing effects arising in the context of obscenity, indecency, and profanity regulation.

[152] I have nothing to contribute to Hart's refutation, which I consider conclusive. Readers interested in these arguments are encouraged to consult the primary sources.

Hate Speech

Hate speech is another context in which the three countervailing effects are very likely to arise.[153] Notwithstanding the Supreme Court's protean positions on the controversy,[154] it seems quite clear that hate speech does possess negligible value and imposes a high social cost.

The hate speech context provides a unique opportunity for observation, because the social sanctions attached to bigoted expressions are especially severe. And the absence of legal sanctions (in the United States) tends to reduce speakers' efforts at concealment. The presence of sanctions and absence of centralized enforcement allow us to easily observe how individuals tend to respond to the imposition of informal speech regulation.

Let us first consider the potency effect. One can hardly fail to observe that the psychic harm arising from racist, sexist, homophobic, or ableist expressions exhibits an inverse relationship to their frequency of use. For example, utterance of the term "nigger" has been abolished from civilized discourse. It is never uttered, but only *referenced* as "the N-word." Due to the infrequency of its expression, the term is imbued with great weight. Its mere utterance, when directed toward a Black individual, rivals the harm of physical violence. When uttered among non-Black individuals, it taken as conclusive evidence of a heinous defect in the speaker.

There can be no doubt as to the loathsomeness of racial bigotry and its effects, and the desire to do *something* to curb its harms is understandable. Yet we should be cognizant of the consequence of reducing the incidence of expressions of hate. Eradicating racial epithets from civilized discourse magnifies their power. It hands to the unapologetic racist, unencumbered by civilized norms, a potent weapon.

[153] In the United States, hate speech has enjoyed relatively strong Constitutional protection. Other putatively liberal democracies have taken a less tolerant stance toward hate speech. *See, e.g.*, §130 of the German criminal code, Strafgesetzbuch, StGB promulgated on 13 November 1998 (Federal Law Gazette I, p. 945, p. 3322) ("(1) Whoever, in a manner that is capable of disturbing the public peace: 1. incites hatred against segments of the population or calls for violent or arbitrary measures against them; or 2. assaults the human dignity of others by insulting, maliciously maligning, or defaming segments of the population, shall be punished with imprisonment from three months to five years."), the French penal code, Code Pénal R. 624-3–4 (prohibiting non-public defamation or insults to individuals on the basis of ethnicity, nationality, race, religion, sex, or sexual orientation), and in the United Kingdom, The Public Order Act 1986 (c 64) §4A, amended by §154 of the Criminal Justice and Public Order Act 1994 ("A person is guilty of an offence if, with intent to cause a person harassment, alarm or distress, he (a) uses threatening, abusive or insulting words or behaviour, or disorderly behaviour, or (b) displays any writing, sign or other visible representation which is threatening, abusive or insulting, thereby causing that or another person harassment, alarm or distress.")

[154] *See supra* §1.4.

This insight is not novel, of course. Discriminated communities have seemingly intuited the danger of the potency effect which arises from the policing of language. Black comedians and musicians have for many decades incorporated use of the term "nigger" into their routines and lyrics. In much the way a vaccine immunizes patients to disease, repeated utterance of the term in innocuous contexts mitigates its potential to hurt in more virulent circumstances. The phenomenon, which activists have termed "reappropriation" or "reclamation," has attracted some scholarly attention, although the research (mainly in sociology and cultural studies departments) wants somewhat for rigor.[155] Analysis in terms of the potency effect represents a possible avenue for improvement.

Reappropriation is a widespread strategy. Although its use by Black Americans furnishes a prominent and easily recognizable example, we can observe a variety of discriminated groups employing it. The disabled community has reappropriated the term "cripple." The Asian American community has attempted to reappropriate "slant."[156] And the gay community seems to have been especially successful in diffusing the pejorative connotations of words like "queer" and "dyke."

Reappropriation is not a recent phenomenon. Deliberate attempts to exploit the potency effect have occurred throughout history. For example, in the eighteenth century, Protestant followers of John and Charles Wesley were pejoratively labeled "methodists." The group embraced that term, accepting it as the proper name of their denomination. So successful was the reappropriation, few people are even aware that the term was once used to denigrate the adherents of that faith.

Let us next consider the inverse supply effect. It follows trivially from the extrinsic value that bigoted speakers derive from causing hurt that the conditions for the inverse supply effect will obtain in the context of hate speech. Given that there is a potency effect—i.e., that decreasing the supply of hateful expressions increases hearers' sensitivity to them—individuals wishing to cause harm will experience stronger incentives to exploit that increased sensitivity. Reducing the supply of hate speech will thus tend to increase incentives for bigots to express it.

Lastly, let us consider filtering. Filtering is the most evident of the three countervailing effects in the context of hate speech. It is practically axiomatic that less racist individuals are more easily deterred from expressing hate speech, and more

[155]More serious scholarship investigating the causes of semantic change exist of course in the study of linguistics. *See* W.V. Quine, Quiddities 53–54 (Belknap Press 1987).

[156]Interestingly, a rock group consisting of Asian members, calling themselves "the Slants," attempted to register for trademark protection, which the Patents and Trademark Office rejected, essentially on the ground that expressions of hate speech would not be granted trademark protection. The case was litigated to the Supreme Court, where musicians prevailed. *See* Matal v. Tam, 582 U.S. _____ (2017).

racist individuals are less easily deterred. There will also be marginal individuals, who are deterred from engaging in hate speech day-to-day, but who deploy the language only on those occasions when they feel it will have the most impact. The increased potency and asymmetric deterrence of hate speech regulation will tend to filter out the low-harm expressions and increase incentives for high-harm expressions.

Fighting Words

Another context in which the three countervailing effects arise is in fighting words. The analysis requires some finesse. The justification for regulating fighting words lies not in the harm caused by speech, but rather in the harm which results *from the harm* caused by speech. The concern is not the psychic injury done to the hearer of fighting words, but rather the consequent public disorder that results when the hearer seeks reprisal through violence. In other words, the primary harm (affront to the hearer) is the catalyst for secondary consequential harms (the public disorder and injury resulting from reprisal). The policy objective is to reduce the secondary harm, and mitigating the magnitude of primary harm—it is supposed—will tend to have the knock-on effect of reducing the secondary harm.

However, regulation of the primary harm is susceptible to the three countervailing effects. Consequently, it is conceivable that regulation would reduce rather than improve social welfare. Consider that if the prevalence of fighting words increases, hearers' sensitivity to the affront will tend to decrease, reducing the probability of violent reprisal. Conversely, if the prevalence of fighting words decreases, then hearers' sensitivity to the affront will tend to increase, raising the probability of violent reprisal. The concern with the knock-on effect does not alter the potency effect analysis.

Likewise, the secondary effects objective, idiosyncratic to fighting words, amplifies the inverse supply effect and filtering. Consider that low-harm injurers are more likely to be deterred by regulation (than high-harm injurers), because their efforts to cause affront are less effective. And high-harm injurers will thus experience increasing incentives to produce fighting words. High-harm injurers may be comparatively more effective at producing fighting words, either because they have a talent for invective, or because are especially good at choosing targets more susceptible to injury.

Moreover, victims are also filtered, because hearers insensitive to fighting words will tend not to find themselves in a circumstance requiring litigation. This leaves only injurers who are especially skilled at insult, and hearers who are especially liable to violent retaliation.

The narrative is a plausible one. A community, where insults and threats are

commonplace, is one where disparaging remarks are more likely to be shrugged off or ignored. An individual accustomed to coarse conversation is less likely to feel the cut of an affronting jibe, and less likely to feel inspired to respond with violence. It is the genteel victim, unfamiliar with harsh treatment, who is liable to feel an obligation to vengeance. So far as the objective is to reduce the incidence of public disorder, the better strategy is to encourage policies which inure the population to affront. The alternative is to cultivate a community of eggshells, astounded by insults to their honor, and inspired to dueling at the slightest provocation.

4.3.2 Persuasive Speech

Let us turn now to the regulation of harms which are supposed to result when hearers are convinced to adopt false beliefs or odious preferences. I will refer to expressions intended to affect hearers' beliefs and preferences as "persuasive speech." In contrast to affronting speech, the harm is not to cause offense, outrage, hurt, or distress in the recipient. Rather, the harm of persuasive speech is in the proliferation "bad ideas."

I include within this category several subcategories, including defamation, commercial speech, and "fake news."

Defamation

Defamation can occur when a group or individual transmits a signal to other individuals about a third party. There are four relevant player categories in the defamation context: (1) the purveyors of true information, (2) the purveyors of false information, (3) the recipients of information, and (4) the subjects of information.

There are two economic justifications for the enforcement of defamation claims. First, that defamatory speech generates an externality. Purveyors of false information enjoy some utility by sending false signals about subjects. And subjects suffer harm to their reputations. Assuming the value (to the purveyor) of transmitting false signals is less than the harm caused (to the subject), the conventional view is that the law should deter the sending of false signals. The argument is that by imposing sanctions on purveyors of false information, the expected benefit (to the purveyor) of purveying false information decreases, and therefore the activity level of purveying false information decreases, effecting a reduction in social cost.

The second justification is that defamation laws have a screening effect. Recipients enjoy utility from the receipt of true information and disutility from the receipt of false information. By imposing a sanction on purveyors of false information (assuming purveyors of false information experience a higher probability of sanction than purveyors of true information), the asymmetric deterrence causes the supply of

false information to decrease faster than the supply of true information (at least up to some socially efficient point), leading to a net improvement in social welfare. In other words, enforcement satisfies the monotone likelihood ratio property, increasing the ratio of purveyors of true information relative to purveyors of false information.

Notice that both justifications depend upon the belief level of recipients. If recipients of false information disbelieve, then neither the recipients nor the subjects suffer any harm. It is only when recipients of false information *believe* the false information that they and the subjects suffer harm. Thus, the problem may be simplified to some extent by focusing our analysis on the belief level of the recipients.

All three countervailing effects are present in the defamation context. Consider the harms suffered by recipients of false information. Observe that recipients' welfare will tend to increase, the greater their belief level in false information, and the weaker their belief level in true information. In other words, if some information X is expressed, and a recipient's belief level is represented by $\alpha \in [0, 1]$, where 0 is complete disbelief, and 1 is complete belief, then the recipient's welfare function U_R will be decreasing as α increases ($\frac{\partial U_R}{\partial \alpha} < 0$) if X is false, and increasing as α increases ($\frac{\partial U_R}{\partial \alpha} > 0$) if X is true. The trouble, of course, is that with any given signal X, recipients will not know whether X is true or false without further investment in search.

If we assume that recipients are rational, and their beliefs are Bayesian, then they will set their belief level according to the prior probability that X is true, and update their belief level to account for the probability that the truth of X produces a confirming signal. In the simplest case, where recipients have no additional information about the relative credibility of a particular purveyor, they will take the probability that a given signal is true to be the supply of true signals σ_T divided by the total supply of signals $\frac{\sigma_T}{\sigma_T + \sigma_F}$ (where σ_F is the supply of false signals). It follows immediately that if the imposition of sanctions decreases the supply of false speech relative to true speech, then the belief level for any given expression will increase. And clearly, there will exist values for which the decrease in the supply of false speech is more than offset by the harm due to the increased belief level.

The intuition here is straightforward. Imagine a world in which everyone defames and no one tells the truth. Clearly, the rational recipient would disbelieve all signals, and the harm caused by false signals would be zero. Now, as the proportion of truth-tellers increases, the credibility of a signal will tend to increase, and the marginal increase in harm for any given instance of defamation will tend to increase as well. Thus, increasing the supply of truth-tellers relative to the supply of defamers will tend to increase the per-unit harm of defamation. In other words, a recipient will tend to rely more when the ratio of purveyors of true information relative to purveyors of

false information is higher.

However, defamation represents a special case. Although the potency effect is clearly present, in the defamation context it cannot result in a net *increase* in harm. If the increasing belief level generated a decrease in the recipient's payoffs, then the *rational* recipient will behave *as if* the belief level were less than their actual belief level. Framed differently, we might disambiguate belief level from "trust level," where the trust level $\beta \in [0, 1]$ is a function of belief level α. Formulated thusly, the rational recipient selects the trust level $\max_\beta U_R(\alpha, \beta)$ given α.[157] It follows trivially that in the limiting case, increasing the sanction level (thereby decreasing the supply of purveyors of false information) will have no effect on the trust level, and thus no effect on the harm generated by the expression of false speech. In case the recipient's trust level is unaffected by a change in the proportion of purveyors of true and false information, *neither* the recipient's harm nor the subject's harm will change. Nevertheless, even when the extremum circumstance does not obtain, the potency effect will act as a friction, reducing the social benefit accrued from a reduction in the supply of false speech. Ergo, the potency effect *can* nullify the effect of changes to the relative composition of purveyors, and it will at least lessen the benefit of reducing the supply of defamers. However, it cannot increase the total harm.

Next, *even if* the imposition of sanctions decreases the relative supply of purveyors of false information, the inverse supply effect will create a further friction on the social benefit of sanctions on defamatory speech. Assuming a prospective defamer's extrinsic value is a function of the potency of his defamatory expressions, it follows that his incentives to defame will increase as recipients become more likely to believe his false statements. Thus, if sanctions are effective in affecting recipients' trust levels, this will effect an increase in the incentives of defamers, creating a further friction on the effectiveness of sanctions.

Finally, *even if* the imposition of sanctions results in a net social benefit, despite the presence of the potency effect and inverse supply effect, there remains a further obstacle in filtering. Suppose there are two types of defamers: low-harm defamers and high-harm defamers. Let us define "low-harm" defamers as being those less credible than "high harm" defamers. In the face of sanctions, prospective low-harm defamers are more likely to refrain from engaging in defamatory speech than high-harm defamers, because they will receive less benefit from transmitting a defamatory signal. The imposition of sanctions will tend to dissuade more low-harm defamers from engaging in speech, and by the inverse supply effect, this will increase the incentives of high-harm defamers to exploit the increased credulousness of recipients. It follows that filtering will further reduce the effectiveness of legal sanctions, and

[157] Assume α is the subjective probability that the signal is true.

in case the change in the composition of defamers (and the difference in magnitude between high-harm and low-harm) is sufficiently large, filtering may even result in a reduction in total social welfare.

Of course, if we assume away information costs, then the rational Bayesian recipient of information will strategically adjust his trust level to account for this as well, ensuring an equilibrium that represents a reduction in total social harm. However, I think the information costs associated with this are likely to be significant, given that the proportion of high-harm and low-harm defamers is likely to be opaque to prospective recipients. And I do not think this should be assumed out of the model. Nevertheless, *even if* we accept this assumption, although the net effect of filtering could not result in a reduction in social welfare, it would again in the extremum case render the imposition of sanctions null, and introduce yet another friction impairing the effectiveness of legal sanctions at a minimum.

Stepping back, the intuition here is easy to grasp. Prospective defamers crave a credulous audience. And the greater the proportion of defamers (relative to truth-tellers), the more skeptical the population of information recipients will be. Thus, allowing defamatory speech—or false statements of fact more generally—to go unregulated will tend to incentivize greater skepticism in the population of hearers, mitigating the harm to the recipients of potentially false information, and also mitigating the harm to subjects of potentially false information. It must be conceded that under some conditions, increased skepticism will reduce the benefit that recipients derive from true information (they will be skeptical of the true information insofar as they are unable to distinguish a true signal from a false signal without an an independent search investment). Yet the point is not that the imposition of legal sanctions will have no effect (or a negative effect) on social welfare, but rather that plausible conditions exist, under which the imposition of legal sanctions *could* have no effect (or a negative effect) on social welfare.

And once again, even if the imposition of legal sanctions in some subset of defamation cases were welfare-improving, their effectiveness will tend to be undermined. And it is then plausible that the costliness of sanctions, litigation costs, and other kickers could tip the balance, such that the regulation of defamatory speech would be net welfare-reducing when all factors are considered.

Commercial Speech

Expressive rights in the realm of commercial speech have historically enjoyed only attenuated protection.[158] In particular, fraudulent inducements and deceptive ad-

[158] *See supra* §1.5.

vertisements are wholly outside the First Amendment guarantee. The argument for the received view is that false signals, designed to entice consumers to purchase a product, service, or property, may reduce incentives to enter into mutually beneficial exchanges. Unprotected from false signals, consumers will forgo many potentially Pareto-improving exchanges to avoid incurring the cost of information gathering required for verification of a signal's truth. Furthermore, false signals can result in consequent harms arising from misplaced reliance on inferior or dangerous goods.

For example, if a tobacco company runs an advertisement claiming that cigarettes whiten teeth, freshen breath, and aid in children's pulmonary development, then it could persuade some individuals to smoke. And if consumers later discovered that smoking actually increases the probability of developing various cancers, heart disease, emphysema, and rancid breath, then they would be reluctant to trust future promises of a product's qualities and effects, refraining from entering into potentially Pareto-improving exchanges. When they do purchase goods, services, or property, they will at least invest more effort in search to validate the claims of advertisements.

Additionally, those individuals who are induced by the false advertisement to take up smoking might become addicted, continue the habit, and succumb to the maladies associated with cigarette smoking. The healthcare expenses which ensue are a further source of social cost, which regulations on deceptive advertising might prevent.

Restated somewhat more precisely, if the promised value of an exchange is B, the cost of disappointment is C, the investment in verification is x, and the probability that the signal is true is $p(x)$ such that $\frac{\partial p}{\partial x} > 0$, then consumers face an expected payoff of $p(x)B-(1-p(x))\,C-x$ in the absence of regulation. Allowing disappointed consumers to collect damages when producers communicate false signals increases the consumer surplus. With regulation, consumers can expect a payoff of B.

The case for regulating fraudulent inducements and deceptive advertising is compelling. However, the cost in a deregulated regime may be somewhat less than the received argument predicts. The freedom to express false signals in a commercial context would not free promisors of their contractual obligations generally. It would simply be an abrogation of the promisee's right to void a fraudulently induced agreement. In order to ensure liability for claims about a product, service, of property, the promisee—knowing he cannot rely on non-promissory claims—will simply insist that the claims be expressed in promissory terms.

Even still, there will be an increase in transaction costs and forgone surplus, and the harm-based reasons to favor regulation persist. This leads us to consider once again the possible countervailing effects. First, with respect to the potency effect, if false signals are regulated, then the ratio of true signals to false signals will tend

to increase. Thus, consumers will decrease investments in verification and increase investments in reliance. This will magnify both consumers' susceptibility to harm and the magnitude of harm when false signals are believed.

Second, with respect to the inverse supply effect, when unscrupulous sellers observe the increasing credulousness of consumers in the presence of regulation, they will want to exploit that gullibility. Additionally, for those products or services where the consumer's reliance entails additional profit—for example, with subscriptions or brand ecosystems—sellers will have even greater incentives to engage in deceptive practices.

Third, with respect to filtering, it will tend to be the producers who sell the most inferior or dangerous goods who experience the strongest incentives to exploit the increased susceptibility of consumers in a regulated regime. Producers with quality products will simply refrain from overstating the excellence of their merchandise. Producers whose goods are only slightly substandard can more easily invest in improving quality or seek buyers who require products of lesser quality with honest advertising. The producer who has the most to gain from exploiting the greater credulousness of consumers is the producer whose goods are in least demand, and for whom it would be most costly to improve.

A full analysis of fraudulent inducement and deceptive advertising would require an excursus well exceeding the scope of this article. Additional complicating factors which a comprehensive investigation might include are: the effect of reputation, insurance, and private information screening services.

The case for regulation seems stronger for these subspecies of commercial speech than in the other speech contexts heretofore discussed. Yet even if the benefit of regulation were sufficient to overcome the countervailing effects, it is nevertheless important that lawmakers be cognizant of the tradeoffs in the design of commercial speech regulation. For example, devising a doctrine analogous to contributory negligence, such that consumers are obliged to undertake reasonable verification in order to claim damages, would help to mitigate underinvestment in search. And a "reasonable reliance" standard would help to mitigate excessive reliance investments.

Fake News

I should like to conclude my inquiry with a phenomenon which has arisen rather more recently: the problem of "fake news." It is becoming ever more apparent that the Russian government in 2016 undertook an active disinformation campaign to bias the U.S. presidential election in favor of Donald Trump. Similar efforts have been observed in the "Brexit" referendum in the United Kingdom and in recent

elections in the Ukraine, France, and Germany. There is little reason to suppose the Russian government will discontinue its activities, and it seems likely that foreign disinformation campaigns are likely to pose an ongoing threat to the democratic process. Indeed, it is a plausible surmise that the effectiveness of the Russian effort will inspire other states to embark upon similar escapades.

Much of the policy discussion concerning Russia's disinformation campaign identifies it as a species of "cyber attack," warranting defensive measures. There is of course an intuitive appeal to the proposition that the one must *defend* when *attacked* by hostile foreign actors. However, we should be wary of a knee-jerk response. Here again, attempts to combat persuasive speech will be likely to result in countervailing effects.

With respect to the potency effect, the analysis here is analogous to that in defamation and commercial speech. If measures aimed at reducing the supply of foreign disinformation are successful, then citizens will tend to invest less effort in search and verification. Assuming citizens *want* to acquire true information—or at least information not motivated by hostile motives—they will naturally tend to undertake greater investment in verification and less investment in reliance when there is a greater risk they are receiving foreign disinformation. Therefore, reducing the supply of foreign disinformation will tend to increase citizens' vulnerability to the residual fake news which remains.

Next, with respect to the inverse supply effect, if the citizenry were made more credulous due to successful efforts to restrict the dissemination of foreign disinformation, then this would render it a yet more enticing target. Hostile foreign actors would be incentivized to redouble their efforts, engaging in more sophisticated and subtle mechanisms to influence voters, made more gullible by the reduction in the supply of fake news. Presumably the payoff from swinging an election would be substantial, and hostile foreign actors could be expected to invest considerable resources to exploit the impressionability of voters.

Finally, filtering may plausibly occur in at least two ways. First, relatively less hostile or less powerful foreign actors would be more likely to be deterred, leaving more hostile or more capable foreign actors to fill the void. A less hostile foreign actor would be more easily deterred, because its interests would be relatively more aligned with the target nation's interests. I take this to be what it *means* to be relatively less hostile. Those less hostile foreign actors would thus enjoy less benefit from interfering in the target nation's democratic processes. Less capable foreign actors would be deterred, because their efforts at disinformation, if not widely received, would tend to be less successful in generating the critical mass of mutually confirming counter-narratives which a comprehensive disinformation campaign would require.

Second, hostile foreign actors may strategically invest in disinformation only when the stakes are highest. Assuming efforts to curb foreign disinformation impose at least some cost upon the purveyors of disinformation, they would tend to focus their resources on efforts calculated to impose the greatest effect. They might save their efforts for those occasions when elections seem likely to be closely fought, or when a potential outcome is anticipated to be especially unfavorable.

It is of course possible that increasingly sophisticated disinformation campaigns may be countered with increasingly sophisticated regulation. However, this commits states to a rent-seeking game. Hostile foreign actors invest in more streamlined disinformation; and target states invest in more savvy regulation. The standoff is liable to lead to ever-escalating investments in expression and suppression, and a considerable dissipation of resources.

It is difficult to say with specificity how well strategies to combat foreign disinformation would fare in the face of the countervailing effects. Target governments are still developing their strategies. It is not yet clear whether such efforts would effect a reduction in supply even *absent* the countervailing effects. As governments work to formulate their responses to this fresh nuisance, I suggest they take seriously the option of *doing nothing*. Admittedly, the initial effect will tend to subvert the democratic process. However, a prevalence of foreign disinformation is, over time, likely to incentivize citizens to invest greater effort in verification, to reduce reliance, and generally to harden themselves against future subversive influences.

I suspect that the Russian government would rather deal with the regulation of social media than a population of skeptics.

4.4 Conclusion

This article reveals that the courts and doctrinal literature have embraced an implicitly economic framework in nearly every controversy concerning the freedom of expression. It is therefore a surprising irony that the economic analysis of speech rights has received so little attention from Law & Economics scholars. This article accepts the latent invitation, undertaking an explicitly economic exploration of the speech right, extending the simplistic harm-deterrence model assumed by courts and prior scholarship to reveal a fuller, more capacious account of the issue.

The received view assumes that the speech realm is an implicit market. If the market analogy is taken seriously, the countervailing effects I identify go to the essence of the speech right and threaten to overwhelm the fundamental justification for speech regulation.

My argument does not entail an absolute prohibition on speech regulation. It merely undermines the logic of the received view. It is a theoretical point, raising the evidentiary burden for the proponent of regulation. More research is surely needed, and I hope some attempt is made to empirically verify or falsify the intuitive narratives I sketch in section 3. It may well be that some of the speech regulations I have discussed are (or would be) efficient law. The final determination whether a regulation is efficient requires empirical supplementation.

If we value the freedom of speech even one half as dearly as our hymns proclaim, it is incumbent upon the law that any diminishment of it be soundly justified. This article shows the defectiveness of the justifications upon which the law has hitherto relied. The mere harmfulness of an utterance cannot support incursions upon its expression.