RATIONAL CHOICE AND CATEGORICAL REASON

BRUCE CHAPMAN[†]

I. RATIONALITY AS A NORMATIVE IDEAL

The theory of rational choice, as understood by most economists and many other social scientists, has both a normative and a positive content. Normatively, it points to what should be done maximally to achieve some given end, and, while it might not prescribe any particular end, it points to what it is to have a consistent set of ends that are capable of being so maximized. For example, if an agent had a set of ends that gave rise to a cyclical ordering of available alternatives, that is, if she preferred *x* to *y*, *y* to *z*, and *z* to *x*, it would not be possible for her to choose any one of these alternatives without another of the alternatives being preferred to it according to her own criteria for choice. In other words, it would not be possible for her to satisfy completely, or maximize, her own ends.[1]

Positively, the theory of rational choice is used to describe, explain, and predict human behavior. Agents are assumed generally to behave in an internally consistent way that can be rationalized by the theory of maximization.[2] Thus, if an agent has already chosen alternative *y* over alternative *z*, and then chooses alternative *x* over alternative *y*, the assumption, and prediction, will be that the agent will choose alternative *x* over alternative *z*.

[†] Faculty of Law, University of Toronto. I am grateful to Jack Knetsch, Rahul Kumar, Arthur Ripstein, and all the participants at the University of Pennsylvania Law School Symposium on Preferences and Rational Choice for their helpful comments on an earlier draft. Research funding from the Social Sciences Research Council of Canada and from the Connaught Fund at the University of Toronto is also gratefully acknowledged.

[1] For a good discussion of rational choice construed as maximization, and the properties that are consequently required for the underlying preference relation, see AMARTYA K. SEN, COLLECTIVE CHOICE AND SOCIAL WELFARE 7-20 (1970).

[2] *See, e.g.*, Amartya Sen, *The Formulation of Rational Choice*, 84 AM. ECON. REV. PROC. 385, 385 (1994) (noting that rationality is assumed to mean acting to maximize a utility payoff); Amartya Sen, *Maximization and the Act of Choice*, 65 ECONOMETRICA 745, 746 (1997) (distinguishing "maximization," which only requires choosing an alternative that is not judged to be worse than any other, from "optimization," which, more strongly, requires choosing an alternative that is better than all others).

Recently, the positive theory has come under attack from experimental psychologists and economists.[3] Their experimental results, gathered together under the banner of behavioral analysis, show that the maximizing model of rational choice often does not provide a very accurate account of how agents actually choose. Moreover, the departures from the model appear systematic rather than random, suggesting that something other than maximization is going on.

However, the general tenor of these studies is not to question the normative ideal of maximization. Rather, the departures from the standard account of rational choice are typically characterized, and criticized, as failures to be rational. Agents are only human beings, after all, and human beings are subject to the limitations that must, inevitably and systematically, arise out of personal biases, limits on the salience and availability of important information, and the distorting effects of how a given problem is framed. Thus, real-world agents are only, it is said, capable of a "bounded rationality," using "rules of thumb" and various "heuristics" (sometimes helpful, sometimes not) rather than the fully fledged maximizing rationality that is still largely accepted as the ideal for rational choice.[4]

In this Article, I argue that for many decision-making problems, the normative account of rationality that animates rational choice theory, and not just the positive account that is criticized by the behaviorists, is deficient, even as a theory of *ideally* rational behavior. Ra-

[3] The literature is now huge. Good selections can be found in CHOICES, VALUES, AND FRAMES (Daniel Kahneman & Amos Tversky eds., 2000); JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES (Daniel Kahneman et al. eds., 1982); RATIONAL CHOICE: THE CONTRAST BETWEEN ECONOMICS AND PSYCHOLOGY (Robin M. Hogarth & Melvin W. Reder eds., 2d ed. 1987). For a wide-ranging textbook treatment of many of the relevant issues, see JONATHAN BARON, THINKING AND DECIDING (3d ed. 2000).

[4] This is quite clearly Jonathan Baron's view. *See* BARON, *supra* note 3, at 66 (noting that, for one reason or another, people often fail to follow prescriptive models of decision making and rationality). However, not all theorists of "bounded rationality" think of its "boundedness" as setting constraints on what, ideally, rationality would otherwise require of us. As one pair of theorists wrote:

> Bounded rationality is, however, not simply a discrepancy between human reasoning and the laws of probability or some form of optimization. Bounded rationality dispenses with the notion of optimization and, usually, with probabilities and utilities as well. It provides an alternative to current norms, not an account that accepts current norms and studies when humans deviate from these norms. Bounded rationality means rethinking the norms as well as studying the actual behavior of minds and institutions.

Gerd Gigerenzer & Reinhard Selten, *Rethinking Rationality*, *in* BOUNDED RATIONALITY: THE ADAPTIVE TOOLBOX 1, 6 (Gerd Gigerenzer & Reinhard Selten eds., 2001). On this view, bounded rationality provides an alternative account of *ideally* rational behavior.

tionality, I shall suggest, provides for an *ordered particularity*, including particular decisions, but the notion of an ordering that informs this alternative account of ideally rational behavior, and which is more appropriate in some decision-making contexts (including many legal ones), is very different from the idea of an ordering that informs the standard account within rational choice theory. The latter, which, as already suggested, is closely allied to the idea of *maximization*, remains largely quantitative and single-minded in its orientation, this despite the pluralism of motivations that it appears to be able and willing to accommodate within its seemingly minimalist structure.[5] The alternative account is more qualitative, or categorical (although not absolute), offering a conception of a rational ordering of particularity that is more allied to the idea of an *understanding* or *interpretation* (under rules or principles) than it is to maximization.[6] At the risk of import-

[5] The structure of conventional rational choice is minimalist in the sense that it only seems to require that an agent (1) be able to order any set of available alternatives from best to worst, and (2) not choose an alternative *x* from that set if there is another alternative that is better than (or more preferred to) *x* according to this ordering. (More structure is required for rational choice over uncertain alternatives, including, most importantly, the so-called "sure thing" principle. For a discussion of this principle, see *infra* note 35 and accompanying text.) Requirement (1) appears to be open to any possible motivation or criterion for choice (including concerns for justice, altruism, respect for the environment, process values, etc.); requirement (2), while capturing the idea of maximization, seems to follow simply from taking these different motivations or criteria seriously. Why settle on choosing some alternative if there is another alternative available that is better according to one's own criteria for choice? However, as I hope to demonstrate in this Article, there is already enough in this apparently minimalist structure to prevent us from accommodating some attractive (nonmaximizing) principles of choice as rational.

[6] Compare the characterization of these two alternative accounts of rational decision making provided by Drazen Prelec:

> Decision analysis, which codifies the rational model, views choice as a fundamentally technical problem of choosing the course of action that maximizes a unidimensional criterion, such as value or utility. The primary mental activity . . . is the reduction of multiple attributes or dimensions to a single one, through a specification of value trade-offs . . . . For rule-governed action, the fundamental decision problem is the quasi-legal one of constructing a satisfying interpretation of the choice situation. The primary mental activity involved in this process is the exploration of analogies and distinctions between the current situation, and other "canonical" choice situations in which a single rule or principle unambiguously applies.

Values and Principles: Some Limitations on Traditional Economic Analysis, *in* SOÇIO-ECONOMICS: TOWARD A NEW SYNTHESIS 131, 131 (Amitai Etzioni & Paul R. Lawrence eds., 1991). For some suggestions about how the differences between maximization and reasoning by analogy might be captured in choice theoretic terms, see Bruce Chapman, *The Rational and the Reasonable: Social Choice Theory and Adjudication*, 61 U. CHI. L. REV. 41 (1994).

ing some unnecessary baggage, but for reasons that I hope will become clearer as the argument unfolds, I refer to this alternative conception of rationality as *categorical reason.* If that phrase suggests a longstanding rationalist tradition, exemplified by Kant, but rejected by the British empiricists like Hobbes and Hume, who are the most likely intellectual ancestors of contemporary rational choice theorists, that is not entirely unwelcome.[7]

The real challenge for this Article, however, is not so much to articulate two alternative accounts of rationality that have had some traditional followers, but to begin to make each accessible to the other within some common intellectual framework. While I think rational choice theory provides a useful and precise set of tools for beginning this process of achieving mutual understanding between the traditions, I shall argue that some quite fundamental postulates of rational choice theory (including some of the axioms of choice consistency and strong independence) will have to be relaxed if the contributions of categorical reason are properly to be accommodated within it. However, I hope to show that there is much advantage in this, even for what the rational choice theorist hopes to achieve, and to illustrate

[7] For a concise account of the intellectual origins of rational choice theory in the works of Hobbes and Hume, see Martin Hollis & Robert Sugden, *Rationality in Action*, 102 MIND (n.s.) 1, 2–7 (1993). In interpretations of Kant, the word "categorical," as in "categorical imperative," is often thought to mean "absolute" or "without qualification"; for an example of this interpretation, see CHARLES FRIED, RIGHT AND WRONG 9-13 (1978). This is not the meaning of "categorical" I mean to suggest in my phrase "categorical reason." *See infra* text following note 77 (discussing the concept of "categorical reason"). Rather, I mean something more like "within categories" or "according to rules," as in the following:

> Everything in nature, both in the lifeless and in the living world, takes place *according to rules*, although we are not always acquainted with these rules . . . . The whole of nature in general is really nothing but a connection of appearances according to rules; and there is *no absence of rules* anywhere. If we believe we have found such a thing, then in this case we can only say that we are not acquainted with the rules.
>
> The exercise of our powers also takes place according to certain rules that we follow, *unconscious* of them at first, until we gradually arrive at cognition of them through experiments and lengthy use of our powers, indeed, until we finally become so familiar with them that it costs us much effort to think them *in abstracto* . . . .
>
> Like all our powers, *the understanding* in particular is bound in its actions to rules, which we can investigate. Indeed, the understanding is to be regarded in general as the source and the faculty for thinking rules in general. For . . . the understanding is the faculty for thinking, i.e., for bringing the representations of the senses under rules.

IMMANUEL KANT, LECTURES ON LOGIC 527 (J. Michael Young ed. & trans., Cambridge Univ. Press 1992) (1800).

the point by reference to some systematic difficulties that the rational choice theorist faces in the theory of social choice and game theory.

Part II reviews the results of some recent behavioral experiments that suggest that agents respond to reasons in a way that is not always consistent with some of the fundamental axioms of (value-based) rational choice. I look at choice involving certain and uncertain alternatives, and focus on the weak axiom of revealed preference in the former and the sure thing principle in the latter. My claim is that while some of the choices that some of these experimental subjects make do seem problematic from a rational point of view, sensible scenarios can be constructed that make good sense of these systematic violations of the rationality axioms.

In Part III, I argue that common law adjudication manifests the same tension between reason-based choice and conventional (value-based) rational choice that was shown in the experiments. However, I argue that the common law idealizes reason-based choice, insisting not only that a claimant be right, but that a claimant be right *and rational*—that is, right for the right reasons. I refer to this reason-based ideal as categorical reason.

In Part IV, I suggest that the idea of categorical reason can be useful both in the theory of social choice and in the theory of noncooperative games. In social choice, categorical reason brings a kind of conceptual discipline to the preferences that can be admitted into social choice, and this helps to avoid certain problems of instability and collective irrationality. In the theory of games, categorical reason publicly organizes the particularity of individual agents' choices so that coordination and cooperation are more likely to occur.

II. RATIONAL CHOICE BEHAVIORISM AND REASON-BASED CHOICE

A. *The Case of Certainty*

One might have thought, or even hoped, that a theory of rational choice would have informed us about how people *think* or *deliberate* about their decisions, or about how their choices are explained or justified by *reasons*. That, typically, is how a legal theorist would understand the obligation to offer an account of rational decision making. At the end of their article on "reasons," for example, John Gardner and Timothy Macklem conclude that rationality "is simply the capacity and propensity to act (think, feel, etc.) only and always for undefeated reasons."[8]

However, the agenda for developments in the economic theory of rational choice has, apparently, been one of psychological reductionism. The idea, which began with Vilfredo Pareto's replacement of cardinal with ordinal utility as a motivation for choice in the early part of the twentieth century,[9] has been to rely less and less on any claims about what might be going on in someone's head.[10] With the advent of revealed preference theory, as originally developed by Paul Samuelson in the 1930s,[11] the expunging of anything psychologically substantial that might explain a set of rational choices, like the maximization of *self-interest*, *utility*, or, now, even *preference*, is more or less complete. What matters for rationality is the *consistency of externally observable behavior*, not any particular subjective motivation.[12] This reliance on what is objectively observable is typically thought to be "scientifically more respectable"[13] than any attempt to speculate about, and model, private thoughts, motivations, or reasons.

Of course, the requirements of a rational consistency of observable choice are not unrelated to the requirements of a rational maximization of unobservable preference or utility. Indeed, the former, while considered a fully autonomous subject matter for the scientific and systematic study of choice, is still typically thought capable of being "rationalized" by the latter. Thus, Samuelson's *Weak Axiom of Revealed Preference* ("WARP"),[14] which is still the central postulate of the new behaviorism in cases of choice over certain outcomes, has been

[8] John Gardner & Timothy Macklem, *Reasons*, *in* THE OXFORD HANDBOOK OF JURISPRUDENCE AND PHILOSOPHY OF LAW 440, 474 (Jules Coleman & Scott Shapiro eds., 2002).

[9] VILFREDO PARETO, MANUAL OF POLITICAL ECONOMY 103-209 (Ann S. Schwier & Alfred N. Page eds., Ann S. Schwier trans., Augustus M. Kelley Publishers 1971) (1927).

[10] For a discussion of the historical developments in the theory of utility as a motivation for rational choice, see MARK BLAUG, ECONOMIC THEORY IN RETROSPECT 343-53 (3d ed. Cambridge Univ. Press 1978) (1962).

[11] P.A. Samuelson, *A Note on the Pure Theory of Consumer's Behaviour*, 5 ECONOMICA (n.s.) 61, 61-71 (1938).

[12] *See, e.g.*, J.R. HICKS, A REVISION OF DEMAND THEORY 6 (1956) ("[T]he econometric theory of demand does study human beings, but only as entities having certain patterns of market behaviour; it makes no claim, no pretence, to be able to see inside their heads.").

[13] *See* I.M.D. Little, *A Reformulation of the Theory of Consumer's Behaviour*, 1 OXFORD ECON. PAPERS (n.s.) 90, 97 (1949) (noting that objective observation has been deemed "scientifically more respectable" because it allows explanation of behavior "without reference to anything other than" behavior).

[14] Samuelson, *supra* note 11, at 62-70.

shown to be logically implied by, and consistent with, what would be chosen by a rational maximizer of preferences.[15] According to WARP, if an agent ever chooses an alternative *x* over alternative *y* from some set of alternatives, then that agent should never (on pain of inconsistency) choose alternative *y* over alternative *x* from any other set of available alternatives.[16]

One is tempted to add "unless, of course, her preferences have changed," but this would be to seek refuge in a preference-theoretic explanation of a possible departure from what is supposed to be a purely choice-theoretic requirement. Nevertheless, the temptation is revealing in that it shows what really lies behind WARP as a plausible requirement for rational choice. The idea, surely, is that an ideally rational agent can arrange all *conceivable* alternatives in order of preference and would choose, from any *available* subset of those alternatives, the one that was highest in that ordering. Such an agent would never violate WARP. Further, an agent satisfying WARP would always choose as if she had such a preference ordering and were maximizing it.

At first glance this last claim might seem odd, since there does not seem to be enough in WARP to generate the thought that there must be an underlying transitive preference relation motivating choice. For example, if an agent chose *x* from the choice set (*x*, *y*), *y* from the choice set (*y*, *z*), and *z* from the choice set (*x*, *z*), there would not yet be any violation of WARP, although the choices do seem to reveal an intransitive preference ordering, something that can frustrate maximization.[17] However, the violation of WARP is manifest if we can require that the agent now show us a consistent choice over the choice set (*x*, *y*, *z*), that is, if we take seriously the idea that the agent must, in a way analogous to the complete preference requirement, be consistent in her choices over any *conceivable* set of available alternatives. For, given her first three choices over the three different pairs, the agent cannot now choose any alternative from the triple (*x*, *y*, *z*) with-

[15] *See* Kenneth J. Arrow, *Rational Choice Functions and Orderings*, 26 ECONOMICA (n.s.) 121, 126 (1959) ("The most interesting conclusion is the complete equivalence of [WARP] with the existence of an ordering from which the choice function can be derived."); *see also* Amartya K. Sen, *Choice Functions and Revealed Preference*, 38 REV. ECON. STUD. 307, 310-11 (1971) (proving the logical equivalence of WARP with choice under a rational preference relation).

[16] Arrow, *supra* note 15, at 123.

[17] *See supra* text accompanying note 1 (introducing the normative theory of rational choice as maximization).

out violating WARP.[18] Thus, a rational maximizer of preferences will choose in a way that satisfies the behavioral requirements of WARP, and an observable chooser satisfying the requirements of WARP will choose as if she had a fully transitive preference ordering that she was maximizing. It does not appear, therefore, that the behaviorist revolution has offered up any real surprises at the level of principle.

Of course, where behaviorism has offered up something new is in the recent experimental research that shows that agents do not actually choose in the way that the most minimal consistency conditions, like WARP, seem to require. That is, the choice between two alternatives *x* and *y* can vary—indeed, it can be reversed—according to what else is in the available set of alternatives. To the extent that this appears to be systematic and predictable, it suggests that these choices cannot be rationalized as the maximization of preference. It is a different question, perhaps, whether they can be rationalized at all.

The most interesting experimental results for the purposes of this Article are those offered by Eldar Shafir, Itamar Simonson, and Amos Tversky around the idea of "reason-based choice."[19] These authors begin by contrasting reason-based choice with value-based choice, the latter being their name for the economic theory of rational choice. On this latter view, a value is associated with each alternative and choice is characterized as the maximization of value. Reason-based choice, on the other hand, is more characteristic of legal scholarship[20] and analyses of historically significant case studies.[21] It "identifies various reasons and arguments that are purported to enter into and influence decision, and explains choice in terms of the balance of reasons for and against the various alternatives."[22] This, they admit, is a somewhat vague characterization of rational choice, and it might not be clear why there would necessarily be any incompatibility between value- and reason-based choice. Surely the "values" of different alter-

[18] *See* AMARTYA SEN, CHOICE, WELFARE AND MEASUREMENT 58 (1982) ("[N]o matter what he chose out of this set of three alternatives . . . he must violate [WARP].").

[19] Eldar Shafir et al., *Reason-Based Choice*, *in* CHOICES, VALUES, AND FRAMES, *supra* note 3, at 597, 597-619; *see also* Amos Tversky & Itamar Simonson, *Context-Dependent Preferences*, 39 MGMT. SCI. 1179, 1179 (1993) (discussing empirical findings inconsistent with value maximization and presenting a context-dependent model of choice).

[20] Shafir et al., *supra* note 19, at 598; *see also infra* Part III (suggesting that common law adjudication is a form of reason-based choice).

[21] For example, reason-based choice analysis has been applied to a study of the Cuban Missile Crisis. GRAHAM T. ALLISON, ESSENCE OF DECISION: EXPLAINING THE CUBAN MISSILE CRISIS (1971), *cited in* Shafir et al., *supra* note 19, at 598.

[22] *Id.*

natives provide good reasons for choosing them, the argument might go, and it seems natural to think that the balance of values would provide a good indication of where the balance of reasons is ultimately to be found. However, their precisely constructed experiments serve to indicate more clearly what is meant by reason-based choice and why the conflict with value-based choice can potentially arise.

The experiments show that agents will often latch onto a reason for choosing a particular alternative just to resolve the conflict that they feel in facing choice. The "irrationality," at least from a rational choice perspective, is that almost any reason, including a seemingly "irrelevant" one, will do. For example, some subjects were asked to choose between alternative *x*, six dollars in cash, and *y*, a high quality pen.[23] The pen was chosen by 36% of the subjects and the remaining 64% chose the cash.[24] However, when the subjects were presented with a choice from these same two options together with a third, *z*, another pen that was of clearly inferior quality to the first, then the percentage of subjects that chose *y*, the higher quality pen, rose dramatically.[25] This appears to suggest that many subjects who would choose *x* over *y*, when only those alternatives are available, will choose *y* over *x* when some third alternative, *z*, is added to the set of available alternatives, a violation of WARP.

The explanation offered is that these subjects now have a *reason* to choose *y*, namely, that it is a pen of clearly superior quality to *z*, a reason that they did not have when *z* was unavailable.[26] However, this pattern of choices does appear somewhat "irrational." The fact that *y* is an alternative that is obviously better than *z* provides a good reason for choosing *y* over *z*, but it appears to provide little reason for choosing *y* over a quite *different* alternative *x*. Indeed, as nothing about the values of *x* and *y* is changed by introducing *z* into the choice set, one might have thought that nothing would change for a rational chooser in a value-based choice between *x* and *y*. This property of "rational" consistency is what is captured by WARP and what is violated so systematically by the experimental results.

Nevertheless, it is not difficult to construct a different scenario where it seems more sensible to think that the addition of some alternative *z* might change the choice between alternatives *x* and *y*. Sup-

[23] *Id.* at 609.

[24] *Id.*

[25] *Id.*

[26] *Id.* at 610.

pose, for example, that someone is offered a choice of fruit at the end of a dinner.[27] If only a large apple, *A*, and a large orange, *O*, are offered to her, she would choose the large apple. Both fruits are large and, all else equal, she prefers apples to oranges. However, if she is offered *A*, *O*, and a small apple, *a*, then different considerations arise. For now there is an issue of etiquette to be addressed. The rule, let us say, is that one should never choose the larger of two items of the same kind. Our chooser now reasons that, in the choice from the set (*A*, *O*, *a*), she cannot now choose *A*, because that would be in breach of the rule of etiquette. She therefore chooses *O*, a piece of fruit that is larger than *a*, but a fruit of a different kind. Thus, from the set of alternatives (*A*, *O*), she chooses *A*; but from the set of alternatives (*A*, *O*, *a*), she chooses *O*, a violation of WARP.

The chooser would also reveal an intransitive preference ordering if the different fruits were offered to her in pairs. She would choose *A* from the pair (*A*, *O*), *O* from the pair (*O*, *a*), and *a* from the pair (*a*, *A*), in violation of transitivity. The reason, of course, is that the rule of etiquette does not come into play until the third choice, when the big and small apples are presented together. Until that point the chooser can select between the fruits purely according to taste, or according to the different values of the different alternatives, choosing the highest valued one; in other words, she can choose in the way that the theory of value-based choice and maximization suggests. But when the two apples are presented together, etiquette becomes an issue between them, that is, as an issue bearing on the *relationship* between those two alternatives, not as a property or value of either of the two alternatives considered on its own. In this way we can say that the concern for etiquette, unlike the concern for taste, is a *partition-dependent* or *categorical* idea; it arises only when the two alternatives, *a* and *A*, appear together within some partition of the alternatives.[28]

[27] This example is now much discussed. The earliest published analysis of it of which I am aware is in Philip Pettit, *Decision Theory and Folk Psychology*, *in* FOUNDATIONS OF DECISION THEORY 147, 163 (Michael Bacharach & Susan Hurley eds., 1991). The example, and close variations of it, is also analyzed in Paul Anand, *The Philosophy of Intransitive Preference*, 103 ECON. J. 337, 344 (1993); Bruce Chapman, *Law, Incommensurability, and Conceptually Sequenced Argument*, 146 U. PA. L. REV. 1487, 1498-99, 1503-05 (1998); Amartya Sen, *Internal Consistency of Choice*, 61 ECONOMETRICA 495, 501 (1993). *See also* Amartya Sen, *Is the Idea of Purely Internal Consistency of Choice Bizarre?*, *in* WORLD, MIND, AND ETHICS: ESSAYS ON THE ETHICAL PHILOSOPHY OF BERNARD WILLIAMS 19, 24 (J.E.J. Altham & Ross Harrison eds., 1995) [hereinafter Sen, *Bizarre*] (showing the epistemic relevance of a different menu of alternatives).

[28] For other examples used to make the same point, see JOHN BROOME, WEIGHING GOODS: EQUALITY, UNCERTAINTY, AND TIME 100-01 (1991); ISAAC LEVI, HARD

The etiquette example provides, therefore, another instance of reason-based choice pulling the chooser in a direction different from that prescribed by the logic of value-based choice. The presence of *a* in the set of alternatives gives our chooser a reason in etiquette for *not* choosing *A*. But, just as for the experimental results referred to earlier, this is a reason that does not seem to be relevant to any comparison between *A* and *O*; it is a very partition-dependent consideration. The values of *A* and *O* as alternatives would appear to be unchanged, and the choice between them, one might have thought, would be unaffected by such an "irrelevant" reason.

Yet, despite this apparent "irrationality," what is happening in the etiquette example is hardly incomprehensible to us, at least if we have any sort of feel for the rule of etiquette that is involved. We simply *understand* the choice situation (*A*, *O*, *a*), where both *A* and *a* are present and etiquette *is* at issue, *differently* from the choice situation (*A*, *O*), where *a* is absent and etiquette is *not* at issue. And this different understanding, which turns on the availability of an alternative that is itself never chosen, requires that a different choice be made over those two alternatives *A* and *O*, which were always available for choice. Thus, it is not as if the different understanding arises simply because a different set of available alternatives means we can now choose something different, and more particularly something *better*, which was not available earlier.[29] That sort of different understanding, which does

CHOICES: DECISION MAKING UNDER UNRESOLVED CONFLICT 33, 105 (1986); James F. Reynolds & David C. Paris, *The Concept of 'Choice' and Arrow's Theorem*, 89 ETHICS 354, 363 (1979).

[29] Not surprisingly, this is how the committed rational choice theorist typically solves the etiquette problem and others like it. *See, e.g.*, BARON, *supra* note 3, at 235 (arguing that, because of the relevance of etiquette, a large apple does not mean the same thing when compared to a large orange as it does when compared to a small apple). What appears as an inconsistent choice over the *same* pair of alternatives is actually a choice over a *different* pair of alternatives and, therefore, the issue of inconsistency cannot arise. Choosing "the big apple *A* from a set where the little apple *a* is available" is simply not the same as choosing "the big apple *A* from a set where the little apple *a* is not available," or so the argument goes. We might even capture this idea by more accurately relabeling the two alternatives for choice as *A*/*a* and *A*/~*a*, respectively (where *A*/*a* is read "*A* when *a* is also available" and *A*/~*a* is read "*A* when *a* is not also available"). Thus, the apparent inconsistency of choosing *A* from the set (*A*, *O*) and *O* from the set (*A*, *O*, *a*), for example, is resolved under this redescription of the problem as choosing *A*/~*a* from the set (*A*/~*a*, *O*) and *O* from the set (*A*/*a*, *O*, *a*), in perfect conformity with choice consistency conditions like WARP. This "solves" the difficulty, but only at the cost of rendering the choice consistency requirements vacuous. As Sen observed, "[i]f every time the set from which the choice is being made changes, the choice of any given alternative . . . is taken to be a different choice[,] . . . then no condition of internal consistency of choices from *different* subsets can make

not affect our understanding of the previously available alternatives, would always be relevant to a maximization of value. Rather, the different understanding arises because the addition of the new alternative changes how the previously available alternatives, themselves unchanged, are now *conceived*. And this new understanding changes how we choose between those previously available alternatives.[30] Thus, this

any demand whatsoever." Sen, *Bizarre*, *supra* note 27, at 26. This is a heavy price to pay to "secure" the conventional requirements of rationality in rational choice theory. Moreover, the very act of redescribing the alternatives according to what else is available in the choice set *concedes* the point at issue, viz., that what we *are* doing in choice, and what we *want* to do under that description, *varies* with the set of available alternatives. This variation is only obscured, and not preserved as a subject requiring more thorough analysis, if we simply provide a new set of partition-*dependent* descriptions of the alternatives to preserve a partition-*independent* choice consistency condition. To his credit, John Broome has at least recognized that the redescription strategy must be refined so that something of the original force of the conventional rationality conditions can be preserved without reducing them to the worst forms of "ad hocery." *See* BROOME, *supra* note 28, at 102-07 (discussing the recognition of rational requirements of indifference as a way of dealing with the problem of emptiness). However, Broome's refinement strategy, which allows alternatives to be "individuated" in the way described earlier, either begs the question (in that rationally justified differences in choice still have to be justified as differences between the alternatives separately considered) or generates a quite different problem for rational choice in that some alternatives cannot now logically be compared, something which violates the completeness requirement. For a discussion of the relationship between completeness and rationality, see *infra* text accompanying notes 80-81.

[30] For an interesting paper relating the conventional choice consistency conditions (like WARP) to the equally conventional monotonicity requirements of classical logic, see RUVIN GEKKER, NONMONOTONIC REASONING AND THE FOUNDATIONS OF RATIONAL CHOICE (European Pub. Choice Soc'y, Working Paper No. 61, 2002), *available at* http://www.economics.nuigalway.ie/downloads/papers/0061paper.pdf. Under monotonicity, if some proposition *p* is sufficient to imply another proposition *q*, then the compound proposition *p and r* should also imply *q*; in other words, the sufficiency of *p* for *q* should not be undermined by adding *r*. However, under non-monotonic or defeasible reasoning, this is precisely what is relaxed. While *p* on its own might be sufficient for *q*, the addition of proposition *r* can imply not-*q*.

In the etiquette example, the non-monotonicity is found in the following: Let proposition *p* be "options *A* and *O* are available for choice," let proposition *q* be "I should choose *A* and not *O*," and let proposition *r* be "option *a* is available for choice." Then, under the obligation to choose something according to the desire to eat (larger pieces of) fruit *and* the rules of etiquette, while *p* implies *q*, *p and r* implies not-*q*. (In fact *p and r* implies, "I should choose *O* and not *A*.")

It is arguable that much of legal reasoning and legal argument proceeds non-monotonically. Certainly the desire to provide an adequate model of legal reasoning has been one of the great motivators for the development of non-monotonic or defeasible logics in recent years. *See, e.g.*, JAAP C. HAGE, REASONING WITH RULES: AN ESSAY ON LEGAL REASONING AND ITS UNDERLYING LOGIC, at xiii (1997) (discussing defeasibility of reasoning with rules, "in particular legal rules"). For example, the addition of a certain legal defense *r*, while not relevant or even admissible as a consideration *until* the prima facie case *p* is in place (it is a *defense* after all), can reverse or defease the leis different choice under a truly different understanding of the (same) alternatives (themselves unchanged in value); it is not different choice simply because (trivially, we now understand that) differently valued alternatives are now available. Where WARP allows the latter role for an altered "understanding," it does not allow the former.

Yet the idea that we might choose differently over empirically indistinguishable alternatives because we have a different understanding of the issues that are at stake in our choice is hardly novel. Consider the example of how one chooses *as a friend*, something Aristotle discussed at some length.[31] It is widely appreciated that there is something problematic about choosing to be someone's friend for instrumental reasons, especially, say, if one is being a friend (or seeking the good for one's friend) simply because it makes one better off. There may be reciprocity or mutual "back scratching" in that, but it fails fundamentally as friendship. The test (although not, of course, the end) of true friendship is in the willingness to continue acting as a friend even if doing so makes one worse off.

This much is elementary, but we can take the basic insight further. A true friend cannot even have the value of having friends as the reason she does what friends do. That is also too instrumental; it puts the value of having friends before the friendship itself. Even if the actor sees her conduct as perfecting friendship, or seeking (too much) to do what it is that friends do, it would still be too calculating and too (self-consciously) goal-oriented for genuine friendship. But suppose she says, in response to some proposal, "That's not what friends do; I cannot do that." Then the concept of friendship *informs* what she does, although it does not *guide* what she does in the way that a goal (e.g., the goal or value of achieving or maintaining friendship) might. She acts under an *understanding* of what friendship is, even of what friendship requires, but she does not, strictly, act that way because friendship requires it. The latter suggests too much that there is a

gal outcome *q* that would otherwise be implied by *p*. I have argued elsewhere that defeasible rules provide an innovative, rational, and peculiarly legal structure for the accommodation of plural values in collective decision making, and one that cannot be captured within the conventions of rational choice theory. Bruce Chapman, *Law Games: Defeasible Rules and Revisable Rationality*, 17 LAW & PHIL. 443, 446 (1998); Chapman, *supra* note 27, at 1494-95; *see also* John L. Pollock, *A Theory of Moral Reasoning*, 96 ETHICS 506, 512-20 (1986) (arguing that defeasibility should provide the logical structure for moral reasoning more generally).

[31] Aristotle devotes two books of *The Nicomachean Ethics* to friendship. *See* John M. Cooper, *Aristotle on Friendship*, *in* ESSAYS ON ARISTOTLE'S ETHICS 301, 301-40 (Amélie Oksenberg Rorty ed., 1980) (discussing Aristotle's writings on friendship).

moment when she can understand the alternatives for choice independent of the concept of friendship, and then go on to choose amongst them as friendship, now brought to bear upon the choice, would have her do. But the concept or category of friendship informs, and orders, the particularity of her choices in a more gapless or immediate way. She sees the choice as directly implicating friendship, as a particular immediately implicates the category of which it is an instance. And she chooses *as a friend*, with the result that, just as immediately, she instantiates the category in the particularity of what she does. In this way, the *concept* of friendship can be the reason for, or the ordering of, her acting the way she does, even if the *value* of friendship cannot.

This is not to say that there is no value to be achieved in friendship or that friendship does not connect to something valuable. If everyone acts under the aspect of friendship, where friendship rationally orders or gives reason to what they do, then the good of friendship is likely to be achieved and enjoyed. And there is value in that. But still, it seems quite plausible to say that the value of friendship is no part of our rationale for acting as friends. We act one way rather than another simply because we know what it is to be friends. The value we achieve is merely an incidental by-product, maybe even an *essentially incidental* by-product,[32] of our acting on this knowledge and for this reason.

This digression into the notion of friendship has served to illustrate that there is a long and durable tradition in the idea that an understanding, or a concept, can act as a reason for choice. This tradition also suggests that there is a difference between reason based on such a concept and reason based on value. The importance of the etiquette example is that it shows precisely where this difference is to be found within the axioms of rational choice theory. Where a different concept gives rise to a different understanding of the alternatives available for choice, even a different understanding of all those alternatives that continue to be available as other alternatives change, then choice may vary according to that changing understanding, *even for*

[32] For a discussion of social states that are "essentially by-products," that is, states that "can only come about as the by-product[s] of actions undertaken for other ends," see JON ELSTER, SOUR GRAPES: STUDIES IN THE SUBVERSION OF RATIONALITY 43 (1983). *See also* Robert Sugden, *Rational Choice: A Survey of Contributions from Economics and Philosophy*, 101 ECON. J. 751, 781 (1991) (linking Elster's idea to problems of self-defeating rationality in the theory of rational choice and, in particular, the problem of rational commitment).

those alternatives that continue to be available. While this might trouble the rational choice theorist committed to certain choice consistency conditions like WARP, conditions that make sense on a view of rational choice that is value-maximizing and goal-oriented, it is less clear that such variations should surprise those whose conception of rational choice is to be found more in the idea that particular decisions are rational to the extent that they can be ordered, or organized, under the aspect of different concepts, understandings, or categories of thought.

B. *The Case of Uncertainty*

The discussion so far has related to choice over certain alternatives. We have been questioning whether the idea of maximizing one's preferences (or values) over those alternatives, even if that only takes its behaviorist form as a choice consistency condition like WARP, is the only sensible conception of rational choice, or whether an alternative conception, sensitive to the different understandings that a chooser might bring to a choice problem, might also provide an account of rational ordering of particular decisions.

However, the most general theory of rational choice, which purports to rationalize behavior as *expected* utility maximization, deals with choice over risky alternatives.[33] Of course, to handle the more general case, some further choice axioms are required. In this Section, we will focus on one in particular, namely, the strong independence assumption or "sure thing principle."[34]

The sure thing principle has been characterized as "[t]he key qualitative property that gives rise to expected utility theory,"[35] and so

[33] Within this more general theory, certain choice is interpreted as the trivial case where the probabilities for the different possible outcomes are reduced to either one or zero.

[34] For a discussion of the axioms underlying expected utility maximization (or the idea that an agent's observable choices over uncertain alternatives can be rationalized, or represented, as the maximization of that agent's expectation of utility), see R. DUNCAN LUCE & HOWARD RAIFFA, GAMES AND DECISIONS 23-31 (1957). Luce and Raiffa refer to the strong independence assumption as "substitutibility." *Id.* at 27. For a discussion of the possible origins of the name "sure thing principle," see LEONARD J. SAVAGE, THE FOUNDATIONS OF STATISTICS 21-24 (1972). "Strong independence" is the name that Paul Samuelson uses for the axiom. *See* Paul A. Samuelson, *Probability, Utility, and the Independence Axiom*, 20 ECONOMETRICA 670, 670 (1952) (asserting that strong independence conditions "create the existence of certain special or canonical indexes of utility and probability that are *additive*").

[35] Amos Tversky & Daniel Kahneman, *Rational Choice and the Framing of Decisions*, *in* CHOICES, VALUES, AND FRAMES, *supra* note 3, at 209, 210.

it is important to have a sense of what it means. Roughly, the principle allows the chooser to cancel or eliminate from her consideration any possible state of the world that yields the same outcome (the "sure thing") regardless of one's choice. Suppose, for example, that one can choose between two lotteries, *A* and *B*, where *A* provides for a seaside vacation if one draws (from 100 possible numbered tickets) a number between 1 and 10 and $100 for any number between 11 and 100, and *B* offers a mountain vacation on the chance that you draw a number between 1 and 10 and $100 for any number between 11 and 100. Since all the possible draws of numbers between 11 and 100 yield the same outcome of $100, the sure thing principle tells us that the choice between the lotteries should turn only on the chooser's preference between the possibilities of vacationing at the seaside or in the mountains.

Suppose that the chooser indicates a preference for lottery *A* over lottery *B*, indicating a preference for the seaside vacation. Now consider a third lottery, *C*, which offers a seaside vacation if one draws a number between 1 and 10 and $200 if one draws a number between 11 and 100. By the same sure thing principle, the choice between lottery *C* and lottery *A* should turn on the chooser's preference between $200 and $100. Suppose, as seems reasonable, that the chooser prefers lottery *C* because, all other (vacation) possibilities equal, it offers (the same chance of) a larger cash award. Thus, *C* is preferred to *A*, *A* is preferred to *B*, and, by transitivity, *C* is preferred to *B*. And surely this last implication (based on transitivity) makes sense even on a direct application of the sure thing principle. If lottery *C* is preferred to lottery *B* in every possible world that might occur (i.e., because of the preference for a seaside vacation over a mountain vacation if a number between 1 and 10 is chosen, and because of the preference for more cash rather than less in the event that a number between 11 and 100 is chosen), then lottery *C* should be preferred to lottery *B* even when the exact state of the world (or which number might be chosen) is still unknown. If this were not the case, that is, if lottery *B* were preferred to lottery *C*, then there would have to be something attractive in the lottery as such, that is, in the particular *combination* of (mutually exclusive) possibilities that *B* offers, namely, a less preferred vacation *and* less cash, which allows lottery *B* to be more attractive for the chooser even though each possibility considered *on its own* is dispreferred. This would, of course, violate the *strong independence* properties of the sure thing principle.

However, despite the apparent rationality of the sure thing principle, behavioral research indicates that it is systematically violated by experimental subjects. The subjects appear to reveal what the behaviorists have called a "disjunction effect."[36] That is, the subjects will indicate a preference for some choice *x* over another choice *y* when they know that event *A* obtains, and they will also indicate this same preference when they know that event *A* does not obtain, but they will indicate a preference for *y* over *x* when it is unknown whether or not *A* obtains. This is a violation of the sure thing principle. What is most interesting about this research is the explanation for why this disjunction effect occurs. As we shall now see, the explanation again points to the importance of reason-based choice.

In one of these experiments, all the student subjects were asked to imagine that they had just taken a very difficult qualifying examination near the end of the fall term.[37] Some of these students were then asked whether they would take advantage of a very attractive holiday package in Hawaii if they knew they had passed the exam.[38] Others were asked whether they would take advantage of the same package if they knew they had failed the exam.[39] A majority of the students indicated that they would choose the vacation package in each of these two possible states of the world.[40] However, when the students were asked if they would choose the package if they did not know whether they had passed or failed, less than a third of the students chose the package and more than 60% were willing to pay five dollars to postpone the decision until the following day when they would know the results of the examination.[41] The fact that students (i) are unwilling to commit to the vacation when there are still two uncertain possibilities before them, namely, pass or fail the examination, but (ii) will choose the vacation in the event that either one of these two possibilities becomes a certainty, illustrates the disjunction effect and violates the sure thing principle.

[36] *See, e.g.*, Shafir et al., *supra* note 19, at 612-13 ("Evidently, a disjunction of different reasons (reward in case of success or consolation in case of failure) is often less compelling than either definite reason alone."); Eldar Shafir & Amos Tversky, *Thinking Through Uncertainty: Nonconsequential Reasoning and Choice*, 24 COGNITIVE PSYCHOL. 449, 449 (1992) (explaining that "[u]ncertain situations may be thought of as disjunctions of possible states").

[37] Shafir et al., *supra* note 19, at 611.

[38] *Id.* at 612.

[39] *Id.*

[40] *Id.*

[41] *Id.* at 611.

The explanation that is offered for this behavior concerns the reasons that the students give for wanting to go on the vacation in each of the possible states of the world. When the student knows that she has passed the exam, she reasons that the vacation is a just reward for her success; when the student fails, she construes the vacation as providing some kind of consolation. Thus, whatever the outcome of the examination, the student has a good reason—albeit a *different* reason for each possible outcome—to take the vacation. However, when the outcome of the examination is unknown, the student lacks either one of these as "a definite reason"[42] for taking the vacation. Thus, under this last scenario, it is as if she knows it is right for her to go on vacation, but not that it is right for any particular reason. This, somehow, makes it hard for her to go.

We could also interpret her difficulty in the following way: when she knows the outcome of the examination, she knows the reason that is relevant to her choice and, therefore, she knows exactly *what it is* that she is doing when she goes on vacation. When she knows she has passed, for example, she knows that this is a "reward trip." Likewise, when she knows that she has failed, she knows that this is a "consolation trip." In other words, she can organize what she is doing under some kind of category or understanding. But when she does not know the results of the examination, it is as if her vacation lacks any such identity, or meaning, for her. Again, it is as if she knows what she should do (after all, she knows that she has no reason to do otherwise), but she lacks any particular understanding of what it is that she is doing when she does it. To the extent that such an understanding might bring order, or rationality, to the particularity of what she does, her reluctance to go, far from manifesting irrationality (as suggested by the violation of the sure thing principle), evidences the importance of this alternative conception of rationality for her behavior. Indeed, without this alternative rationality to bring order to what she does, she will pay five dollars to wait the one day so that she will know what it is that she is doing. And she will do this despite the fact that waiting the extra day will not change what she will do.

The vacation example suggests the importance for choice of having an understanding of what it is that one is doing, as opposed to having no such understanding at all (or, at least, an ambiguous or equivocal understanding because of uncertainty). However, other experiments reveal that the presence of uncertainty, giving rise to a dis-

[42] *Id.* at 612.

junction effect, can generate, more positively, an *alternative* understanding to that arising from choice under certainty and, therefore, give rise to choices that violate the sure thing principle because there is this different understanding informing choice. Consider, for example, what Eldar Shafir and Amos Tversky unearthed about how subjects play the familiar two person prisoner's dilemma game.[43] In their experiment the rate of cooperation in the game was 3% when the subjects knew that the other player had defected, and 16% when they knew that the other player had cooperated.[44] Now one might well have expected some rate of cooperation between 3% and 16% when the subjects were uncertain whether the other player had cooperated or not. However, when the subjects were confronted with this uncertain situation, the rate of cooperation rose significantly to 37%,[45] a number that cannot be explained as some weighted average between the strategy "cooperate when the other cooperates" and the strategy "do not cooperate when the other does not."

Shafir and Tversky attribute this pattern of responses, a pattern that shows the disjunction effect and violates the sure thing principle, to the different understandings that a subject will have of her choice situation depending on whether she knows if the other player has already made his choice of strategy.[46] When she knows that the other player has already chosen his strategy, whether it be to cooperate or not to cooperate, the subject thinks of herself as acting "on her own."[47] *Given* the choice of the other player, she alone will determine the final outcome of the game. This encourages her to bring a highly individualistic perspective to bear on her choice of strategy, a perspective that leads her more naturally to choose against cooperation. However, in the disjunctive or uncertain situation, all four possible cells of the prisoner's dilemma game are still very much in play, with the outcome of the game still to be determined by a combination of the strategy choices of both players taken together. Shafir and Tversky argue that this provides for a more collective understanding of the situation, and from this more collective point of view the optimal

[43] Shafir & Tversky, *supra* note 36, at 452-59.

[44] *Id.* at 454-55. The higher rate of cooperation in the latter situation provides some support for the idea that players will sometimes reciprocate cooperation from the other player rather than always choose the dominant noncooperation strategy that rational choice theory typically prescribes.

[45] *Id.* at 455.

[46] *Id.* at 457.

[47] *Id.*

strategy for both parties is to cooperate.[48] Thus, it is less surprising that cooperation is chosen more frequently in this situation, say Shafir and Tversky, the sure thing principle notwithstanding.[49]

It is worth emphasizing, again, that these differences in understanding that the subject brings to her choice situation cannot be explained on the basis of some variation in the properties or values of any one (or more) of the alternative outcomes considered each on their own. Nor are they to be explained by variations in properties or values of the outcomes as weighted by the probability of their occurrence. Rather, the differences in understanding arise because the same alternatives are somehow "differently conceived" depending on what else is available, something that goes to a *relationship* that exists between alternatives in the set of possibilities. For example, in the disjunctive case of the prisoner's dilemma, the presence of all four cells together as possible outcomes of the game helps to frame any one of them as instances of the "still-to-be-collectively-determined" category. On the other hand, when only two cells within a given column of the game are available (indicating that the other player has already chosen his strategy), the two possible outcomes become instances of the "it's-all-up-to-me" category. The idea that there could be a relationship between alternative outcomes, related as instances of a particularly conceived *category* of possibilities, and that this could influence an individual's choice of a strategy producing those outcomes, is an idea that the independence properties in the sure thing principle are set to deny.[50]

[48] *Id.*

[49] *Id.* at 457-58. Someone might object that there is a better explanation for the subjects' tendency to choose "less rationally" (in the sense that rational choice would require that the subject choose the dominant noncooperative strategy) in the case where the four possible cells of the prisoner's dilemma are still in play. The argument might be that choosing over four cells is a more complex choice than choosing over two cells, and that it should not be surprising that subjects do less well, or less "rationally," in the more complex case. However, Rachel Croson has experimental results that show this cannot be the explanation; it appears that the disjunction effect disappears in games that are equally complex but which do not have the same scope for different reasons. *See* Rachel T.A. Croson, *The Disjunction Effect and Reason-Based Choice in Games*, 80 ORG'L BEHAV. & HUM. DECISION PROCESSES 118, 129-31 (1999) (testing an alternative theory of decision making, complexity, as an explanation of the disjunction effect). Thus, Croson concludes that the explanation for the effect is reason-based, not complexity-based. *Id.* at 131.

[50] The famous Allais paradox can be thought of in the same way. Maurice Allais, *The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School*, *in* EXPECTED UTILITY HYPOTHESES AND THE ALLAIS PARADOX 27 (Maurice Allais & Ole Hagen eds., 1979). Allais argued that agents often

III. REASON-BASED CHOICE AND THE LAW

The common law is more than a mere collection of authoritative resolutions for disputes. In addition to providing a decision, common law judges are typically expected to articulate a set of publicly comprehensible reasons in support of the decision. Indeed, the obligation to provide reasons for one's choices may well be the one thing that effectively distinguishes the common law as a method of collective decision making. In market or political forms of decision making, for example, individual rights holders and legislators can make perfectly authoritative decisions without good reason.[51] But, in the common law, the reasons that are publicly articulated in support of some decision form a large part of the authoritative basis for it. Weak reasoning will undermine the authority of a case and leave it exposed to the indignity of being distinguished into oblivion, if not completely overturned or reversed.

This suggests that common law adjudication is very self-consciously a form of reason-based choice. It remains to be seen, however, whether we can find the same tension in law between reason-based choice and value-based choice that we saw above in the behaviorists' experiments. In this Part, I will argue that the same tension does exist, and that it is often properly resolved in favor of legal reasons that, for groups of judges, can be sensibly organized under particular categories of understanding, the sort of thing that explained the behavior of the experimental subjects discussed earlier. This rational ordering of particular decision making is something I call categorical reason.

will have a "certainty preference" that cannot be allowed for if the strong independence condition, or sure thing principle, of expected utility theory is assumed. The certainty of getting some given payoff in *every* possible state of nature is not a property of any one state of nature. Rather, it is a property of, or a property of how we think of, all the (mutually exclusive) alternative states taken as a whole, that is, a property of their relationship. Moreover, because the alternative states are mutually exclusive possibilities, their relationship cannot be causal, only conceptual.

[51] I do not mean to suggest that the passage of legislation is never informed by reasoned debate. Indeed, for the "republican" theory of politics, this sort of deliberative exchange is often thought to be central and important for grounding the authority of political decision making in general. I only mean to argue that the authority of any given legislative act is, ultimately, a matter of its proper positing by some legitimate lawmaker (e.g., a majority of the legislators). It is not affected by the possibility that, after reasoned debate, the particular reasons that prevailed were not especially good.

To see the connection between law and reason-based choice as understood by the behaviorists, consider the following example:[52] Suppose that a plaintiff was injured while using some product and that she has advanced two separate and independent claims for the recovery of damages from the defendant manufacturer. The plaintiff has argued that the product was either defectively manufactured (*D*) or sold with an inadequate warning (*W*), where *either* of these two arguments would be sufficient, if successful, to win a judgment (*J*) for the damages in question. In symbols, (*D or W*)→*J*. Now suppose that a panel of three judges has considered the various arguments and that each judge has come out on the issues in the following way. Judge *A*

[52] The example is an instance, in disjunctive form, of what Lewis Kornhauser and Lawrence Sager originally called "the doctrinal paradox." Lewis A. Kornhauser & Lawrence G. Sager, *The One and the Many: Adjudication in Collegial Courts*, 81 CAL. L. REV. 1, 3, 57 (1993); *see also* Lewis A. Kornhauser & Lawrence G. Sager, *Unpacking the Court*, 96 YALE L.J. 82, 114-15 (1986) (illustrating the problem of the doctrinal paradox); Lewis A. Kornhauser & Lawrence G. Sager, Group Choice in Paradoxical Cases 2 (Nov. 26, 2001) [hereinafter Kornhauser & Sager, Group Choice] (unpublished manuscript, on file with author) (defining paradoxical cases as those "where a group's consensus on underlying premises diverges from that group's consensus on the ultimate outcome"). For other discussions of the doctrinal paradox, including analyses that link it to more general problems in the theory of social choice and the theory of games, see Geoffrey Brennan, *Collective Coherence?*, 21 INT'L REV. L. & ECON. 197, 200-01 (2001) (examining the doctrinal paradox and illustrating how justice may conflict with doctrine); Bruce Chapman, *More Easily Done Than Said: Rules, Reasons and Rational Social Choice*, 18 OXFORD J. LEGAL STUD. 293, 312-18 (1998) [hereinafter Chapman, *More Easily Done*] (analyzing reasons, legal issues, and structure-based equilibrium); Bruce Chapman, *Rational Aggregation*, 1 POL., PHIL. & ECON. 337, 341-50 (2002) [hereinafter Chapman, *Rational Aggregation*] (arguing that the aggregation of reasoned individual judgments is less subject to paradox than the aggregation of individual preferences); Bruce Chapman, *Rationally Transparent Social Interactions*, *in* COGNITION, RATIONALITY, AND INSTITUTIONS 189, 197-98 (Manfred E. Streit et al. eds., 2000) (using a legal example to illustrate the point that some strategic interactions are transparent simply because they are more rationally comprehensible); Christian List & Philip Pettit, *Aggregating Sets of Judgments: An Impossibility Result*, 18 ECON. & PHIL. 89, 96-100 (2002) (proving a general impossibility theorem for the aggregation of individual judgments); Christian List & Philip Pettit, *Aggregating Sets of Judgments: Two Impossibility Results Compared*, SYNTHESE (forthcoming 2003) (manuscript at 10-12, on file with author) (comparing Arrow's impossibility theorem with their own impossibility result), *available at* http://socpol.anu.edu.au/pdf-files/W20.pdf; Philip Pettit, *Deliberative Democracy and the Discursive Dilemma*, *in* SOCIAL, POLITICAL, AND LEGAL PHILOSOPHY 268, 274-76 (Philosophical Issues vol. 11, Ernest Sosa & Enrique Villanueva eds., 2001); Philip Pettit, Groups with Minds of Their Own 1-4 (Apr. 2001) (unpublished manuscript, on file with author) (discussing the doctrinal paradox in the context of purposive groups more generally), *available at* http://socpol.anu.edu.au/pdf-files/W12.pdf. Saul Levmore used the specific example in the text, involving two different causes of action, to make a quite different point about the conjunction of probabilities. *See* Saul Levmore, *Conjunction and Aggregation*, 99 MICH. L. REV. 723, 729 n.11 (2001) (identifying a "reverse conjunction" problem in the context of the product rule combining independent probabilities).

believes that there is no reason for thinking that the product was defectively manufactured. Thus, she believes argument *D* to be false. But she thinks argument *W* is true; in other words, she thinks that, although the product has not been defectively manufactured, there is some risk and the consumer has been inadequately warned. She would find reason for *J* on the basis of *W*. Judge *B*, on the other hand, believes that, while adequate information has been provided and, therefore, that argument *W* is false, the product has, nevertheless, been defectively manufactured and, therefore, that argument *D* is true. Thus, she too would find in favor of *J*, albeit for a reason different from Judge *A*. Finally, Judge *C* rejects both arguments *D* and *W* as false and, therefore, also rejects *J* as false. The views of the three different judges are presented in summary form in Table 1.

Table 1

	1. There should be *J* for reason *D*.	**2. There should be *J* for reason *W*.**	**3. There should be *J* for reason *D* or *W*.**
Judge *A*	False	True	True
Judge *B*	True	False	True
Judge *C*	False	False	False
Majority	False	False	True

Where:
J is the proposition, "The plaintiff wins a judgment for damages."
D is the proposition, "The product was defectively manufactured."
W is the proposition, "There was an inadequate warning of product risks."
And: (*D or W*)→*J*.

It should be apparent that, at the group level, there is something odd going on here. A majority of the court rejects proposition 1, "There should be *J* for reason *D*," as false. Further, a majority of the court also rejects proposition 2, "There should be *J* for reason *W*," as false. Yet, a majority of the court accepts the disjunctive proposition 3, "There should be *J* for reason *D or W*," as true. There is something collectively irrational in this combination of votes. At the group level, it looks like we have generated a disjunction effect similar to what we observed earlier in the behaviorist experiments.

The collective irrationality would become particularly apparent if the majority of judges who support an outcome favoring the plaintiff, *A* and *B*, had to articulate a set of publicly comprehensible reasons in support of their decision. There are, after all, only two possible arguments that the plaintiff can make in this case to win a reward for damages *J*. Yet, on each of these essential arguments, the two judges who form a majority in favor of the plaintiff have completely opposing views (as indicated by columns 1 and 2). They would, therefore, have some difficulty *saying together* what it is that they want to *do together*. The two judges may share a common preference, or predisposition, for the outcome favoring the plaintiff, but it is not at all clear that they have a shared understanding of what it is that they are doing to reach that outcome.[53] It is as if what they want to do together lacks any real categorical *identity*. Nor is the problem merely that there is a plurality of reasons here, with no particular reason commanding majority support. There *are* majority views on each of the relevant reasons and they are that neither *D* nor *W* is a good reason for the plaintiff to prevail in this action.

Now one might ask whether this lack of any common understanding of, or reason for, a shared preference for a particular legal outcome must translate into any practical difficulty for the majority actually to act on its preference, that is, whether the failure to be able to articulate a common sense of what it is that one is doing should impact at all on the possibility of rational choice.[54] In some strictly causal

[53] Of course, *each* judge has a reason for what she wants to do. To that extent, the position of each judge is reasoned and not merely a matter of preference or predisposition. But as these reasons are not *shared* between the judges, what is shared, not being supported by reason, looks more like a brute preference or predisposition.

[54] If the different members of a majority are voting in favor of a given outcome, but each for different reasons, then in an important sense they are not actually voting on the same proposition. For example, in Table 1, column 3, while it appears that Judges *A* and *B* agree in their voting, Judge *A* is actually voting for the proposition "*J* for reason *W*," whereas Judge *B* is voting for the proposition "*J* for reason *D*." This severely complicates the epistemic support that majority voting gives to a given proposition by way of Condorcet's famous jury theorem. Condorcet's theorem shows that if each member of a group of voters is more likely to be right than wrong about some given proposition (and this probability of being right is more or less equal for each voter), then a majority of such voters is even more likely to be right than wrong about that proposition than is any of the voters alone. MARQUIS DE CONDORCET, ESSAY ON THE APPLICATION OF MATHEMATICS TO THE THEORY OF DECISION-MAKING (Paris, L'imprimerie Royale 1785), *reprinted in* CONDORCET: SELECTED WRITINGS 33, 48-49 (Keith Michael Baker ed., 1976); *see* Bernard Grofman et al., *Thirteen Theorems in Search*

sense, of course, there cannot be any such impact. It is always possible to pursue one's preferences without good reasons, and possible for a majority to pursue its preferences without any coherent reason across its members. The preferred alternative is not less available as an opportunity for choice simply because the majority cannot organize its thinking in favor of that preferred result under a single (coherent) set of concepts or categories of thought. Thus, at first glance there is little in this legal example, perhaps, that provides any reason for thinking that there is some necessary connection between what we can say or think (together) and what we can do (together). The conventions of rational choice theory, therefore, which emphasize (as WARP does) consistency of choice over the consistent availability of unchanging alternatives, seem not to be much affected, although one might have hoped that the idea that one can pursue one's preferences without good reasons would give a *rational* decision theorist some reason to pause.

However, the experimental results provided by the behaviorists[55] do suggest that subjects will have some difficulty making a choice (or making the same choice) if they cannot organize what they want to do under a set of particular (or the same set of particular) reasons. Thus, this literature supports the idea that there can be a genuine tension between what we want to do and what we have reason to do, and that this tension is often resolved, as a matter of fact, in favor of reason.

Further, it seems that our legal practices might provide some normative support for resolving the tension in this way. For suppose the plaintiff in the Table 1 example were to argue that she should win a judgment *J* because there appears to be a majority agreement in favor of this outcome in column 3. In other words, she points simply and bluntly to the fact that a majority of the judges thinks that she should win, albeit for different reasons. A lawyer is tempted to reply, I think, that the plaintiff's appeal to the column 3 agreement is inade-

of the Truth, 15 THEORY & DECISION 261, 264-65 (1983) (formalizing and extending Condorcet's original jury theorem). The theorem provides some reason for using majority voting in legal decision making to get at matters that have a correct legal answer, or are questions of judgment rather than preference, but it is crucial for the application of the theorem that the relevant majorities have voted for the same, or at least not inconsistent, propositions. Thus, in the context of the doctrinal paradox, this means that the Condorcet theorem may provide more epistemic support for the majority votes down columns 1 and 2 than it does for the majority vote down column 3. For further discussion of the relevance of the doctrinal paradox to the application of the Condorcet theorem, see Chapman, *Rational Aggregation*, *supra* note 52, at 341-44.

[55] *See supra* text accompanying notes 37-49 (discussing the results of Shafir, Simonson, and Tversky concerning the idea of "reason-based choice").

quate because in law she has an obligation to frame her claim against the defendant as an *argument*, that is, under some sort of conceptual structure. It will not do for the plaintiff to show only that a majority of the judges believes (in some unstructured way) that the defendant owes (or, even, probably owes) her damages for *some* reason. Instead, she must show that (more probably than not) a majority of the judges believes that there is *a* reason—at least one *particular* reason, here woven out of some particular account of transactional wrong—for the claim. It is in this respect that some claims, even if they are right (or right more probably than not), may not be right for the right reasons. We might reasonably wonder, therefore, whether they are both right *and rational*.

IV. CATEGORICAL REASON AND RATIONAL CHOICE

The argument so far has shown that there is an alternative conception of ideally rational choice, which I have called categorical reason, that requires us to relax some quite fundamental axioms of the economic theory of rational choice. If agents choose under an understanding of what it is that they are doing, and not merely in a goal- or value-oriented way, then they will choose differently under different understandings, the requirements of WARP and the sure thing principle notwithstanding. I have called this alternative conception of rational choice categorical reason because it emphasizes the idea that a relationship between certain alternatives, or how they are understood together or as a category, can inform choice. In this respect, categorical reason is to be contrasted with axioms like WARP and the sure thing principle, which emphasize the independent properties and values of particular alternatives, and insist on choice consistency so long as these properties, independently considered (either within a set of certain alternatives or within a lottery of uncertain outcomes), remain the same.

Further, the research done by the behaviorists has suggested that, as a matter of fact, this alternative ideal of rational choice does inform how agents actually choose. Of course, the particular experiments might still have us wondering how rational it is merely to react (unthinkingly) to some of these different understandings of one's choice situation. For example, to choose a higher quality pen over some cash merely because a lesser quality pen has been added to the set of alternatives[56] does not seem to be a particularly rational thing to do on any plausible account of rationality. However, I have argued that different scenarios can easily be constructed that make perfect sense of this sort of behavior and, further, that the practice of common law adjudication, in its traditional emphasis on the obligation to offer a particular argument in support of one's claim, idealizes a form of categorical reason.

In this Part, I argue that there is some advantage in this alternative conception of rationality, even for what the economic theorist of rational choice seeks to accomplish. I focus on two areas of rational choice theory that have confronted systematic difficulties, namely, the theory of social choice and the theory of games, to suggest how the arguments might go. In the theory of social choice I suggest that the idea of categorical reason can bring some conceptual discipline to bear on the individual preferences that are deemed admissible into the social choice rule and that certain problems of instability in social choice can thereby be avoided. I relate this argument to certain formal results in social choice theory dealing with required restrictions on the domain of admissible preferences. In the theory of games, I argue that categorical reason allows a player to conceive of the game she is playing in a way that makes certain strategy choices less accessible to her. The result, I suggest, is a greater propensity for players to coordinate their choices in a coordination game and, more controversially, a greater propensity for them to choose cooperatively in the prisoner's dilemma game.

A. *Categorical Reason and Social Choice*

In many situations calling for collective action, it seems likely that the individual members of a decisive majority will not have reasons in common for what they most want to do. Yet we will not feel that this is in any way problematic for them.[57] Consider, for example, that most mundane of economic transactions, the purchase of a car. To give this problem a collective dimension, imagine that there is a three-person purchasing consortium and that a majority of the consortium has voted to buy a white sports car. One member of the majority has voted this way only because the car is white and the other only because it is a sports car. Table 2 summarizes this scenario in a way that ap-

[56] This choice is discussed *supra* text accompanying notes 25-26.

[57] *See, e.g.*, Kornhauser & Sager, Group Choice, *supra* note 52, at 18-20 (discussing the different normative premises that might motivate individual members of a group).

pears to make it fully analogous to the earlier legal example laid out in Table 1.

Table 2

	1. We should buy this car because it is white.	**2. We should buy this car because it is a sports car.**	**3. We should buy this car because it is a white sports car.**
Individual *A*	True	False	True
Individual *B*	False	True	True
Individual *C*	False	False	False
Majority	False	False	True

The fact that the members *A* and *B* of the majority in column 3 do not have "reasons in common" to support their shared preference for buying this particular car is not thought to present them with any real difficulty. Nor is it thought to be rationally compelling that this purchasing consortium chooses *not* to buy this white sports car simply because a majority rejects both the idea of buying it because it is white and the idea of buying it because it is a sports car. The decision to purchase a car, even (more particularly) a white sports car, is not essentially decomposable into two prior atomic propositions: Is it a sports car? Is it white? That underlying structure, while possibly a helpful guide to the purchasing decision, is not an essential part of the problem in the same way that the plaintiff's claim to damages in Table 1 needs to be grounded in a particular account of transactional wrong. Rather, the purchasing consortium is out to purchase a car, perhaps even the *best* car that is available to it, *all things considered.* But that judgment is ultimately made *of the car* and *on the whole,* not on a criterion-by-criterion (or column-by-column) basis.

Despite this structural difference in the examples, there might be something useful, even for what the economist seeks to accomplish by way of social choice or what a purchasing consortium seeks to achieve in the market for cars, in insisting on the greater rationality requirement that is inherent in the legal idea that members of a group can act *sensibly* together only if they can organize what they would prefer to do under a common understanding, that is, only if they can act together under a common set of categories or concepts. That this is sometimes difficult to do, and that it sometimes frustrates the achievement of shared preferences, might be precisely what is so useful about it.

To illustrate this point, suppose that the three individuals in our Table 2 purchasing consortium, considering the joint purchase of a car, originally had preferences over three alternative cars as follows (where for each individual the alternatives are preferred in order from top to bottom within each column):

Table 3

Individual *A*	**Individual *B***	**Individual *C***
White sports car (*WS*)	Black sports car (*BS*)	Black family car (*BF*)
Black family car (*BF*)	White sports car (*WS*)	Black sports car (*BS*)
Black sports car (*BS*)	Black family car (*BF*)	White sports car (*WS*)

This is, of course, the preference profile that makes for the familiar majority voting paradox.[58] A majority prefers *WS* to *BF*, *BF* to *BS*, and *BS* to *WS*. Thus, within the social choice framework, there is the danger here of a kind of *excess* of rational doing: for every alternative that one is tempted to choose, there is another that a majority would prefer to have instead. It is this excess of rational doing that gives rise to cycling and instability.

Now it is common for economists to point out that the problem here is that individual preferences are not "single peaked"; there is no general agreement (1) that all the alternatives are to be assessed according to some single decisive dimension, and (2) that one of the alternatives is of intermediate value on that decisive dimension. If only that were so, the argument goes, then that intermediately placed alternative would never be the worst alternative for any voter and the majority voting paradox would be avoided.[59]

[58] For a good introductory discussion of the majority voting paradox, see DENNIS L. MUELLER, PUBLIC CHOICE II, at 63-66 (1989).

[59] *Id.* at 64-66. If there is this sort of agreement across the voters, and preferences are single peaked, then the alternative chosen under majority rule will be the one which possesses that amount of the decisive dimension which is most preferred by the median voter—that is, the voter who is in the middle of the distribution of voters ordered along the decisive dimension according to where their most preferred alternative is on that decisive dimension. Why this alternative would be chosen, and why no other alternative would defeat it under majority rule, can easily be appreciated in the following way: Imagine beginning at the extreme left (or right) of the decisive contin-

This is, in effect, to insist that individuals organize their preferences in a single-minded way along one decisive dimension and to allow them only the limited scope of ordering the social alternatives according to how these alternatives vary *quantitatively* (more or less) along that decisive dimension.[60] But, as the example suggests, and as multidimensional models show more generally,[61] individuals react, reasonably, to a broad range of *categorically* different dimensions or aspects of the social alternatives on offer. And so the question arises whether these different and plural dimensions of a social choice problem can be rationally organized in some way so that instability can be avoided.

The car example is suggestive. The majority coalition of *A* and *C* can *say together* (in support of what they might *do together*), "Given that the car is black, we would prefer it to be a family car." Likewise, the majority coalition of *B* and *C* might be able to say, "Given that it's a sports car, we would prefer it to be black." In this respect, these coalitions can make use of what are sometimes referred to as *generic* preferences.[62] But what would the majority coalition *A* and *B* say together? In some sense, of course, they have a shared preference over a pair of very particularly described alternatives just like the other majority coalitions do. Indeed, as already intimated, that is what gives rise to the instability. But their shared preference for *WS* over *BF* lacks any of the generic structure that characterizes the shared preferences of the other two majority coalitions. Thus, it is harder for them to articulate their shared preference in any sort of categorical way, that is, in a way that makes use of the generic preferences that are in play in the choice problem. In this respect they, as a coalition, are rendered

uum. Then any move to the right (or left) will receive the support of a majority of the voters until we arrive at the median voter's ideal position on the continuum, at which point any further move to the right (or left) will be defeated by a majority of the voters.

[60] The standard example is the ordering of candidates for political office from "left" to "right" on the ideological spectrum. Another example might be the different quantities of some uni-dimensional public good that different voters want to buy at a given tax price. For a good discussion of both of these examples, see ALLAN FELDMAN, WELFARE ECONOMICS AND SOCIAL CHOICE THEORY 169-70 (1980).

[61] *See* Richard D. McKelvey, *Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control*, 12 J. ECON. THEORY 472, 472-82 (1976) (detailing how multidimensional voting can "end up at any other point in the space of alternatives"). For a discussion of the multidimensional case in a legal context, see Chapman, *More Easily Done*, *supra* note 52, at 300-16.

[62] *See* Jon Doyle & Richmond H. Thomason, *Background to Qualitative Decision Theory*, AI MAG., Summer 1999, at 55, 58 (defining generic preferences as preferences among classes).

"speechless," just like judges *A* and *B* were in the Table 1 legal example. But now we can see that there may be some stabilizing effect in using the discipline of a shared (or public) categorical reason to restrict the formation of this majority group. After all, without this additional discipline and structure, there is only a senseless (i.e., noncategorical, nonconceptual) aggregation of (merely particular) preferences and the cycling problem that this permits.

The discipline that is provided by a (public) categorical reason can be related more generally to a particular form of "value restriction" (specifically, "not-between value restriction") that Amartya Sen has shown is sufficient for avoiding the majority voting paradox.[63] Specifically, if all individuals agree that in *any* triple a given alternative is "not between" the other two, that is, is either best or worst of the three, then the majority voting paradox cannot occur. For instance, in the car example, it is easy to see that an alternative way to express what *A* and *C* have in common is their view that *WS* is a "not-between" alternative for them; the real issue between them is whether the purchased car should be a black car (*BF*, *BS*) or not (*WS*). Individual *A* puts the white car alternative (*WS*) first and the pair of black car alternatives (*BF*, *BS*) last, whereas individual *C* has the reverse ordering of these two partitions. This is something that they could have decided first, before they went on to decide, if necessary, what was a secondary issue to them, viz., what type of car a black car should be.[64] Likewise, what the coalition of *B* and *C* has in common might have been expressed as an agreement over *BF* as a "not-between" alternative, the sort of agreement that asks each to decide first whether the car chosen should be a sports car or not and, second, if it should be a sports car, whether it should be black or white. But, again, the pair of individuals *A* and *B* would have some difficulty *articulating* its own version of a common understanding of the relevant issues in this way. They agree between them that *BS* is a "not-between" alternative, but

[63] Amartya K. Sen, *A Possibility Theorem on Majority Decisions*, 34 ECONOMETRICA 491, 492-95 (1966).

[64] It could be, of course, that *A* feels there is a great deal more at stake in the choice of car type than the choice of color, viz., that the preferential distance between *BF* and *BS* is large compared to the preferential distance between *WS* and *BF*. But this cannot be a view that she has *in common* with *C*. For *C* the preferential distance between *BF* and *BS* is *contained within* the distance between *BF* and *WS*. So the search for a *shared* (or *public*) *categorical* reason for choice, at least one linking *A* and *C*, cannot be found here. As the text following this note suggests, this suggested interpretation (that car type is a more important issue than car color) is better for the pair of voters *B* and *C*.

what, exactly, is the category or concept that embraces the partition (*WS*, *BF*) of alternatives that is the complement to that not-between alternative? The problem, again, is that it is hard to "make sense" of such a partition of the alternatives in terms of the categories or concepts (color and type of car) that are in play in the example. We might say, as we can for all the other possible pairs of individuals, that individuals *A* and *B* agree at the level of preferences, but that they do not share any sort of categorical agreement about the sorts of issues that inform their choice and the order in which these issues might be considered.

Now one might object that the imposition of a categorical discipline on preferences still leaves too much unresolved to be helpful. After all, even those two groups of voters, *AC* and *BC*, which (unlike group *AB*) agree that the salient issues are the type and color of the car to be purchased, disagree fundamentally on the *order* in which these two issues should be addressed. For *AC* the most salient issue is color; only after considering that issue would this group turn its attention, if necessary, to what type of car it should be. But for the coalition *BC* the most important issue is type of car, and only if a sports car is chosen would the coalition turn its attention to the issue of color. Moreover, the order in which the issues are considered is likely to affect the outcome; in this respect the matter is analogous to the problem of path-dependent choice.[65] For example, if color is considered first, then it seems less likely that *BS* will end up being chosen. Indi-

[65] The choice of an alternative is path-dependent if the probability of that alternative being chosen varies with the order in which it is presented for consideration as compared to other alternatives. In social choice theory, the conventional view is to think of path dependence as a kind of arbitrariness to be avoided; alternatives should be chosen, the argument goes, according to the value of their intrinsic properties. Kenneth Arrow, for example, defended his use of a collective rationality condition in social choice on the ground that collective rationality, in the form of a fully transitive social preference relation, would guarantee path independence. KENNETH J. ARROW, SOCIAL CHOICE AND INDIVIDUAL VALUES 120 (2d ed. 1963); *see also* Charles R. Plott, *Path Independence, Rationality, and Social Choice*, 41 ECONOMETRICA 1075, 1075-91 (1973) (discussing the relationship between path dependence and collective rationality). However, not all path dependence should be construed as *arbitrary* path dependence; some choice sequences, or paths, might make "more sense" than others and will often matter a great deal to a process theorist. *See* Bruce Chapman, *Individual Rights and Collective Rationality: Some Implications for Economic Analysis of Law*, 10 HOFSTRA L. REV. 455, 466-70 (1982) (reconciling path dependence with deontological thought); Chapman, *More Easily Done*, *supra* note 52, at 303 (arguing that "some legal choice paths or processes are so permeated by thought," and so conceptually privileged, that they make choosing on alternative choice paths almost "unthinkable"); Bruce Chapman, *Rights as Constraints: Nozick Versus Sen*, 15 THEORY & DECISION 1, 2-8 (1983) (discussing the implications of rightful choice partitions for collective rationality).

vidual *C* will vote in favor of black cars, and individual *A* against black cars, in the first round. Whether black cars are chosen categorically in that round depends a good deal on how individual *B*, whose preferences are not categorical in this way because they do *not* satisfy not-between value restriction on alternative *WS*, actually votes. But, in the event of a first round vote for a black car, it does seem likely that *BF* will defeat *BS* in the vote on the issue of type of car. An analogous argument would suggest that *WS* is the less likely choice if the issue of type of car is decided first.

However, in some contexts, there is good reason to think that this sort of path dependence will be less of a problem for categorically sensitive choice.[66] This is because the categories or concepts that make sense of certain *partitions* of the alternatives for choice will often make sense of certain *paths* (or sequences of those partitions) as well, at least if we want to continue to make use of the stabilizing effects of not-between value restriction. To see this, consider the example of a criminal trial, where the two issues to be decided are the verdict and the sentence for the accused. Again, one could imagine a panel of judges considering three possible final outcomes—innocent (*I*), guilty with a severe sentence (*GS*), and guilty with a lenient sentence (*GL*). And again, a natural partition of the alternative outcomes might be into the two issues, verdict (generating the choice "(*I*) or (*GS*, *GL*)?") and sentence (generating the choice "*GS* or *GL*?"), a partitioning that would "make sense" in a way that the alternative partitions, "(*GS*) or (*I*, *GL*)?" or "(*GL*) or (*I*, *GS*)?," would not. (What single concept, category, or issue sensibly comprehends the partition (*I*, *GS*), for example?) But, still, it seems that one could take these two issues, and the partitions to which they lend sense, in order of either "sentence first, verdict afterwards" or "verdict first, sentence afterwards." The law adopts the second of the two possibilities (and the Queen at Alice's trial in Wonderland adopts the first),[67] but is there any reason to do so? One answer, of course, is simple economy: why bother attending to the issue of sentencing until we know that the verdict decision makes it necessary? But the analysis provided here suggests a different answer. While both sequences respect the *partition* of the alternatives that makes the most sense, only the *path* that has us consider the ver-

[66] *See* Chapman, *supra* note 27, at 1507 (arguing that law exemplifies a "categorical application of plural considerations to decisionmaking" by using a "process of adjudication as sequenced argument").

[67] LEWIS CARROLL, ALICE'S ADVENTURES IN WONDERLAND 187 (Univ. Microfilms 1966) (1865).

dict first, or the one forcing the initial choice to be over the partitions "(*I*) or (*GS*, *GL*)," imposes any sort of not-between value restriction on the panel of judges. Under the verdict first sequence, each judge must order her preferences around the salient legal categories, deciding whether to put the alternative *I* either better or worse than (but not between) the alternatives *GS* or *GL*. The sentence first sequence, on the other hand, while paying a kind of lip service to the same set of issues, does not require any of the judges to order her preferences around those issues. For example, a judge who preferred the three alternatives in the order *GL* first, then *I*, and then *GS*, that is, someone who might be saying, "Whether or not I would find him guilty of the offense depends on the sentence he would receive," would have no difficulty voting these preferences under the sentence first procedure even though these preferences do not seem to show a categorical commitment to the issues that are salient in the case. The verdict first sequence, on the other hand, *does* force this judge to ask a more categorical sort of question about the verdict, that is, to show the same sort of commitment to the issues in the case as does the law she personifies.[68] Furthermore, under a verdict first procedure, we not only make sense of the issues in the case, but we also impose a domain restriction on the preferences that legal decision makers can bring to bear on legal decisions so that certain problems of instability are avoided.

The burden of this Section has been to show that categorical reason can provide a useful sort of conceptual discipline on the kinds of shared individual preferences that should be decisive in social choice. Certain individuals, who want to do something together, might find that it is more difficult to act as a decisive coalition in favor of their shared preferences if they are obliged to think, and talk, about exactly

[68] Requiring this sort of structure can, of course, tempt the judge to "nullify" a possible guilty verdict for fear of risking the worst (for her) possible sentencing outcome *GS*. Verdict nullification has attracted a good deal of critical comment, particularly in the United States where, in jury trials, there is the possibility of the death penalty. Juries are said to be charged with the responsibility of reaching a verdict within the law as explained by the trial judge; it is the task of the judge to determine the sentence. For members of the jury to worry about the sentence rather than the verdict, particularly if they think the accused has committed the offense in question, is thought by some to violate the rule of law. Whatever the merits of verdict nullification by juries, my analysis here, based on the stabilizing impact of imposing the categorical constraints of not-between value restriction, offers an independent reason for supporting the verdict first procedure. *See generally* Darryl K. Brown, *Jury Nullification Within the Rule of Law*, 81 MINN. L. REV. 1149, 1155 (1997) (examining jury nullification and suggesting that jury nullification can "occur *within* the rule of law, rather than subvert it").

what it is that they are doing. To the extent that the problems of instability that arise in social choice are explained in large part because too many decisive coalitions can form too easily, the conceptual discipline that categorical reason provides in this respect could be very helpful. I have also tried to relate the idea of categorical reason to some well-known results in the theory of social choice that impose restrictions on the domain of individual preferences that can be admitted into social choice if instability is to be avoided. While the latter results are not in any way new for the economic theory of social choice, it is novel to motivate these results in, and connect them to, the more philosophical idea of categorical reason. I now will suggest that categorical reason can also have a beneficial impact on the possibility of coordination and cooperation in the theory of games.

B. *Categorical Reason in Noncooperative Games*

Consider the simple two person pure coordination game called "Heads and Tails."[69] Each person, without consulting the other, must turn up either "Heads" or "Tails" on her own coin. If each person turns up "Tails"—a match—then each will win five dollars from the pot. However, if each turns up "Heads"—another match—then each will win ten dollars from the pot. In the absence of a match, each wins nothing. What should each person do? What is the *rational* thing to do?

Note that there is no conflict of interest in this game. The two players will receive identical payoffs in all four possible outcomes and, therefore, order these four outcomes in an identical way.[70] Specifically, they both agree that the outcome generated by each of them playing "Heads" is best, that the outcome wherein each plays "Tails" is

[69] This game is discussed in Sugden, *supra* note 32, at 775.

[70] Thus, this could be a game in which all the players are act utilitarians: each seeks to act in such a way that total welfare for her society is maximized, but must do so without the benefit of prior consultation about what she should do to achieve that shared goal. Such a group has a coordination problem (the inability to communicate is what makes the game "noncooperative"), even though there is an identity of interests across the players. *See* D.H. HODGSON, CONSEQUENCES OF UTILITARIANISM: A STUDY IN NORMATIVE ETHICS AND LEGAL THEORY 58-62 (1967) (illustrating that even correct application of act utilitarianism would not necessarily have better consequences, and would possibly have worse consequences, than would acceptance of specific conventional moral rules and personal rules); DONALD REGAN, UTILITARIANISM AND CO-OPERATION 66 (1980) (analyzing the problem of coordination between act utilitarians). For a good introductory discussion showing how rational choice theory implies counterintuitive results in the analysis of coordination games, see Sugden, *supra* note 32, at 774-78.

second best, and that the two non-matching outcomes, "Heads-Tails" and "Tails-Heads," are tied for worst. One might have thought that this would make the choice of actions relatively easy: each player would choose that action, "Heads," which so clearly, and without conflict, makes both players better off.

Surprisingly, however, the choice of this action is less obvious for a player deemed to be rational in the way that rationality is understood within game theory. This is because rational play for any one player depends crucially on what that player thinks the other player will do in the game. It is simply false, the argument goes, to think that one should always turn up "Heads." If the other player turns up "Heads," then, but only then, should the first player match with "Heads" herself. Otherwise, the first player should turn up "Tails" and secure the second best of the matching outcomes. The problem, of course, is that both players are thinking through this same problem of strategic choice at the same time (or, at least, without prior consultation or revelation of their choices), and so neither can really condition her choice on the given choice of the other. Moreover, that each player is rational in this way is typically assumed to be common knowledge in the game.[71] Thus, each player knows that the other is likewise attempting to work out this conditional strategy which conditions on a strategy that is itself conditional on the strategy of the first. The result is an infinite (self-referential) regress that has the effect of leaving each player in a kind of strategic limbo, unsure about what to do.

Nor do the difficulties disappear if we allow the individual player to develop a strategy that appears to recognize and confront this problem as one of uncertainty. The mixed (or probabilistic) strategy that survives the common knowledge assumption requires that each player play "Heads" with a 1/3 probability and "Tails" with a 2/3 probability.[72] However, while this allows both players simultaneously to step

[71] Common knowledge is information which is known to all the players in a game, which each player knows the others know, which each knows the others know that she knows, and so on. Common knowledge of rationality (and of the rules and payoffs of the game) is typically crucial for solving games because it allows players to put themselves in the place of other players, to replicate their reasoning (that is, think through what they will rationally do in their situation), and act accordingly. For a discussion of the importance of the common knowledge assumption to game theory, see CRISTINA BICCHIERI, RATIONALITY AND COORDINATION 39-43 (1993); Chapman, *supra* note 30, at 443-45.

[72] This is the Nash equilibrium mixed strategy. "A Nash equilibrium is an array of strategies, one for each player, such that no player has an incentive (in terms of improving his own payoff) to deviate from his part of the strategy array." DAVID M. KREPS, GAME THEORY AND ECONOMIC MODELLING 28 (1990). Any other assignment of

out of the strategic limbo in which they originally found themselves, the consequence is hardly comforting. Now, under this choice of a mixed strategy by each player, the most preferred outcome where each player matches on "Heads" arises with only a 1/9 probability (the product of each player independently playing heads with a 1/3 probability). One might have hoped that rational choice would do better than that.

The economist Michael Bacharach has characterized the sort of thinking that generates this difficulty as thinking in an "I/he" frame.[73] The "I/he" frame accommodates the idea, central to game theory and Nash-like thinking, that a player should ask what strategy is best for herself *given* what the other player might do, and allows that player, again under common knowledge of such reasoning, to replicate that

probabilities between the two choices, "Heads" and "Tails," would not be stable as each of the two players tested out its rationality under common knowledge, an assumption that allows each to replicate the reasoning of the other and then make corresponding adjustments in a proposed strategy choice.

For an explanation of how to derive (and interpret) a mixed strategy, see ERIC RASMUSEN, GAMES AND INFORMATION: AN INTRODUCTION TO GAME THEORY 69-73 (1989). Mixed strategies require an odd interpretation. The idea is to calculate the probability distribution over one's own possible actions such that the other player will be indifferent between which strategy she chooses. Thus, in the game "Heads and Tails," if player 1 chooses to play "Heads" with a 1/3 probability, the expected payoff for player 2 in playing "Heads" is equal to the expected payoff in playing "Tails," namely, 10/3. But given her indifference between playing either of the two pure strategies (i.e., either "Heads" or "Tails" with certainty) in such circumstances, she should also be indifferent between playing either of those pure strategies and any mixed strategy which combines them probabilistically, including her Nash equilibrium strategy that has her playing "Heads" with a 1/3 probability. Thus, we can say, given that player 1 plays her Nash equilibrium strategy of "Heads with a 1/3 probability," player 2 has no incentive to deviate from her Nash equilibrium strategy of "Heads with a 1/3 probability" since she does no better for herself by so deviating. (The fact that she also does no worse is a problem for the theory in that it is essentially an *equilibrium* theory rather than a theory for how to play the game *ab initio*; why, of all those strategies over which she is indifferent, does she feel any compulsion to play the Nash equilibrium strategy in particular? For an indication of how one game theorist handles this problem, see Robert J. Aumann, *Correlated Equilibrium as an Expression of Bayesian Rationality*, 55 ECONOMETRICA 1 (1987). Aumann describes a correlated equilibrium approach that does away with the dichotomy usually perceived between the "Bayesian" and the "game-theoretic" view of the world by synthesizing the two viewpoints and consequently not requiring explicit randomization on the part of the players. *Id.* at 1.) And we can also say all this of player 1 if player 2 chooses to play (her Nash equilibrium strategy) "Heads with a 1/3 probability." Thus, the playing of "Heads" with a 1/3 probability by each player is a Nash equilibrium for the game since no player, given the strategy choice by the other player, can improve her own payoff by adopting an alternative strategy.

[73] Michael Bacharach, "We" Equilibria: A Variable Frame Theory of Cooperation 5 (June 24, 1997) (unpublished manuscript, on file with author).

same sort of thinking in the other player as well. The other sort of thinking that Bacharach identifies is thinking in a "we" frame.[74] The "we" frame encourages each player to think about the *profile* S of strategies (one for each player) that should be adopted by the players as a group and then identifies the rational strategy for each player as the one that simply (categorically, nonconditionally) has that player "doing her part" S_i within that overall profile.[75] Unlike in the "I/he" frame, a player in the "we" frame does not have to consider whether the other players are themselves doing their parts as components of this profile of strategies in order to justify her strategy choice. Rather, in response to any question about *why* she was doing what she was doing, she would only say, "This is simply what *we* do when we do S (as best)," or, perhaps (to emphasize how the collective understanding orders the particularity of her individual choice), "This is simply what *I* do when *we* do S (as best)," or even, most provocatively (because most categorical in tone), "This is simply *what it is* for us, you and me, to do S (as best)."

It should be apparent that Bacharach's "we" frame is closely akin to the collective understanding that Shafir and Tversky propose as an explanation for the disjunction effect that they observed in the play of the prisoner's dilemma game.[76] It will be recalled that there was a greater propensity for an experimental subject to cooperate when the strategy choice of the other player in the game was still uncertain. In that situation, the outcome of the game still had to be collectively determined by the strategy choices of both players, something that put each player in a more collective (and, it seems, a more cooperative) frame of mind. On the other hand, when the strategy of the other player is *given*, be it to cooperate or not, then the game becomes one in which the one remaining player chooses to determine the outcome of the game, something that provides for a more individualistic frame of mind. This experiment essentially reproduces the "we" frame and

[74] *Id.*

[75] Robert Sugden's notion of "team reasoning" has a similar structure. *See* Robert Sugden, *Team Preferences*, 16 ECON. & PHIL. 175, 176 (2000) (arguing that "the theory of choice *should* allow 'teams' of individuals to be decision-making agents and *should* allow such teams to have preferences"); Robert Sugden, *Thinking as a Team: Towards an Explanation of Nonselfish Behavior*, 10 SOC. PHIL. & PUB. POL'Y 69, 89 (1993) (asserting that "reasons for cooperating do indeed exist, but that these reasons can get a grip only if we conceive of ourselves as members of a team").

[76] *See supra* text accompanying notes 37-45 (noting that the disjunction effect that occurs depends on whether a player knows if the other player has already chosen her strategy).

the "I/he" frame for the subjects, and provides some empirical support for Bacharach's dualistic account of thinking.

It is important to emphasize that what Bacharach's account provides for, and what the Shafir-Tversky experiments support, is the idea of categorical reason, or the notion that it is a different *understanding* that informs choice under the "we" frame. It would be a mistake, for example, to think that the "we" frame only introduces a different, more collective, sort of *motivation*, one that merely identifies "doing one's part" with the (more conditional, less categorical) idea that "I will cooperate if she does." To begin, the latter idea is not consistent with what the Shafir-Tversky experiments show; the subjects tended to be noncooperative almost as frequently when the other player was known to be cooperating as when the other player was known to be not cooperating. Second, this sort of conditional cooperation would do nothing to get the players out of the strategic limbo of the pure coordination game, a limbo that arises precisely because of an infinite regress of mutually conditioning conditionals. Third, the idea of doing one's part as (merely) conditional cooperation would have no impact on the play of the prisoner's dilemma game, where, in game theory at least, the player has a dominant (not a conditional) strategy not to cooperate *regardless* of what the other player does. Rather, what the Bacharach account provides, and what the experiments support, is an idea powerful and categorical enough to take us beyond the problematic regress of the pure coordination game and as far as thinking, at least presumptively,[77] that what the other player does, and what one should do *given* what the other player does, is not even the right way to think about strategic choice. The last thought undermines dominance thinking in the prisoner's dilemma as much as it circumvents the infinite regress of the pure coordination game.

However, now the worry might be that we have ended up with an account of rational cooperation that is too *unconditional*, that is, one that is implausible precisely because it ignores what the other player might be doing. Indeed, this is what might be suggested by the very word *categorical*, and the somewhat Kantian overtones in the phrase *categorical reason*. The Kantian, it is often said, cooperates absolutely, or just because it is right, and without regard to the contingencies of what others might choose to do. But, however Kantian that might be,

[77] This is an important qualification, already hinted at *supra* note 7. The idea is to allow coordination and cooperation to get started, not to commit to either *absolutely*.

it is a mistaken understanding of what is meant here by categorical in the phrase "categorical reason."

Consider again the earlier etiquette example.[78] That example involved categorical reason because the issue of etiquette was only *relevant* for those partitions or sets that included the big and small apples as alternatives for choice. For the other partitions or sets, etiquette was *not relevant at all.* In that limited partition-dependent sense, a sense problematic for WARP, the concern for etiquette was categorical. But it would be a mistake to think that the concern for etiquette was *absolute* as compared to, say, the hedonistic interest in eating larger pieces of fruit. It might be, therefore, that the difference in size between the large and small apple could become so large that the hedonistic interest in larger pieces of fruit would overwhelm (even rightly) the concern for etiquette. In such a situation, with the partition-dependent effect overwhelmed, there would indeed be transitivity of preference and no violation of WARP. However, the point of the example was not to argue that transitivity or WARP *never* obtain, but only to suggest that these properties need not *always* obtain in the way that rational choice theory suggests. Thus, in this respect, the extreme or uncompromising view is the one offered by rational choice theory, not the one offered by the theory of categorical reason.

And the same could be said for the theory of cooperation based on categorical reason. An agent acting under a collective understanding or "we" frame might begin presumptively and *categorically* with the thought that she should "do her part" in *S* because that is what it is for us, you and me, to do *S* (as best). But the agent need not think of herself as *absolutely* committed to cooperation under strategy *S*. If too few others do their part, for example, there may be no "whole" of which one's own individual choice can sensibly be construed or understood as a part. This may call for a rational revision of what it is that one is doing and allow, therefore, for the possibility of not cooperating if others are not cooperating as well.[79] However, this should

[78] *Supra* text accompanying notes 27-28.

[79] For a discussion of conditional or presumptive cooperation as a kind of "revisable rationality," see Chapman, *supra* note 30, at 472-76. I have argued elsewhere, Bruce Chapman, *Rational Voluntarism and the Charitable Sector*, *in* BETWEEN STATE AND MARKET: ESSAYS ON CHARITIES LAW AND POLICY IN CANADA 127 (Jim Phillips et al. eds., 2001), that this account of presumptive cooperation provides a better explanation of voluntary contributions to public goods, such as in the relief of poverty through charitable contributions, than does the theory of rational choice (never cooperate) or Kantian obligation (always cooperate). Moreover, the presumptive cooperation account can make more sense of the tax treatment of charitable contributions, and the empiri-

not be thought of as resurrecting the idea of a *purely* conditional cooperation. A purely conditional cooperation is still too much in the "I/he" frame, and makes no sense at all of a (prior, albeit only presumptive) collective understanding of one's action. What categorical reason rationally requires, therefore, is a defeasible presumption in favor of cooperation, not an absolute (and thoughtless) commitment to it.

CONCLUSION

There is a kind of "incompleteness" in the idea of categorical reason that should freely be admitted. An agent who only sees, or understands, alternatives for choice under the aspect of more general concepts, or categories of thought, does not, perhaps, fully appreciate these alternatives in all their particularity. Any given categorization need not be crude, of course, but short of reproducing a range of categories that is as detailed and fine as the particular alternatives it seeks to organize, it seems inevitable that something must be lost if choice is to be ordered by categorical reason.[80]

In rational choice theory, by contrast, the fully rational agent can compare all possible alternatives for choice. It is true that agents in the actual world are not thought to be fully rational in this way, but that is the ideal. Thus, when we say of someone, "She bought the Volvo because she likes durable cars," in rational choice theory we mean to concede that she probably approached the problem of choice as best she could, but also that, ideally, she would not have limited herself by these broad generalizations and would have compared (in detail) all the possible alternatives that she might have chosen. Rough categorizations and broad rules of thumb are only needed because an agent must make her way through what would otherwise be an "incomprehensibly large number of alternatives, most of which represent unimportant variations on each other."[81]

cal evidence on how individuals respond to these different tax incentives, than can rational choice or Kantian theory. *See id.* at 130 (outlining the complex motivational structure of homo socioeconomicus).

[80] The incompleteness of choice ordered by categorical reason can easily be appreciated if one reconsiders the criminal trial example, *supra* text accompanying notes 67-68. While all three alternatives discussed there are available as a final choice, the "verdict first, sentence afterwards" choice process (which makes outcome *I* the "not-between" alternative in the triple) does not permit a (sensible) pairwise comparison (or choice) between alternative *I* and either alternative *GS* or alternative *GL*. This noncomparability of certain pairs of alternatives violates completeness.

[81] Doyle & Thomason, *supra* note 62, at 61.

For the rational choice theorist, therefore, something of rationality is lost as we move from a fully particular comparison of all possible alternatives to a comparison constrained by categorization. But there must be something gained for rationality as well. For what makes a comparison of all possible alternatives in their full particularity "incomprehensible" is not merely that the set of alternatives is "large." It is also that, without *some* such categorization, the particularity of choice would literally be "unthinkable." We think through to the particulars of our world, after all, only under the aspect of more general concepts or categories of thought.

So we should not be surprised that there is a notion of ideal rationality that competes with the full rationality of rational choice theory and pulls us in an opposite direction, that is, from particular to general rather than from general to particular. However, I hope to have shown in this Article that the tendency to reduce the general to the particular, all in the name of a more fully rational choice, continues to plague rational choice theory. Sometimes this tendency shows up as the temptation to see alternatives for choice in a partition-independent way, as if features of the alternatives themselves were all that mattered for choice and never features shared with other alternatives in the choice set. This is what our discussion of the choice consistency condition WARP revealed. At other times, the propensity for particularity is manifested in the tendency to reduce what is attractive in a whole to what is attractive in its parts. But, as we saw in our discussion of the sure thing principle (or strong independence condition), our understanding of a choice situation and, therefore, what we should rationally do under that understanding, varies according to whether the choice is seen as a whole or as a disjunction of its parts.

These different notions of rationality have been in play in the behaviorists' experiments on choice for some years now. But, for the most part, the results of these experiments have not been organized under an alternative conception of rational choice. This Article has tried to suggest that the alternative conception that is required, one based on categorical reason, is part of a long-standing theoretical tradition, and that rational choice theorists would do well to look to this tradition to solve some systematic difficulties that they confront in social choice theory and the theory of games. That law forms part of this tradition of categorical reason suggests further, perhaps, that legal theorists have a special obligation to show them the way.