

Rationalization*

Vadim Cherepanov, Timothy Feddersen and Alvaro Sandroni.

September 21, 2009

Abstract

In 1908 the Welsh neurologist and psychoanalyst Ernest Jones described human beings as *rationalizers* whose behavior is governed by "the necessity of providing an explanation." We construct a formal model of rationalization. In our model a decision maker selects the best feasible alternative (according to her preferences) from among those that she can rationalize. We show that this theory can accommodate several behavioral anomalies and yet the theory is falsifiable and can be tested non-parametrically like the standard theory of choice. Rationalization theory produces a formal way to use (perhaps contradictory) speech to reveal agents' motivations and a formal way to interpret experimental evidence where revealed preferences and observed choice need not coincide. In particular, rationalization theory can be used to reveal hidden preferences for discrimination. In addition, rationalization theory can be used to understand behavioral changes in the absence of changes in preferences, incentives and opportunity. Finally, rationalization theory can be easily incorporated into game theory.

*We thank Eddie Dekel, Jennifer Jordan, Paola Manzini, Marco Mariotti, Scott Presti, Marciano Siniscalchi for useful comments. Vadim Cherepanov [vadimch@sas.upenn.edu] and Alvaro Sandroni [sandroni@sas.upenn.edu] may be reached at the Department of Economics, University of Pennsylvania. Tim Feddersen [tfed@kellogg.northwestern.edu] may be reached at the Kellogg School of Management, Northwestern University.

1. Introduction

In 1908 the Welsh neurologist and psychoanalyst Ernest Jones wrote a paper entitled “Rationalisation in Every-day Life.” Jones writes: “[e]veryone feels that as a rational creature he must be able to give a connected, logical and continuous account of himself, his conduct and opinions, and all his mental processes are unconsciously manipulated and revised to that end.” While Jones credits Freud with the critical insight “that a number of mental processes owe their origin to causes unknown to and unsuspected by the individual” his paper provides a careful definition of the process of *rationalization*—“*the necessity of providing an explanation.*”

The idea of rationalization has become so well accepted that pundits write about it in the popular press.¹ Psychologists emphasize the facility with which people create explanations for their behavior. While some explanations may be colorful, the need to rationalize does not have an economic impact unless, on occasion, the inability to rationalize constrains economically relevant behavior.

While economics easily accommodates observable constraints an agent’s ability to rationalize a particular choice may be difficult to observe. Such *psychological constraints* have not yet received much attention in the academic economics literature. An exception is Roth (2007) who lists several examples of potentially beneficial practices that were deemed repugnant and banned. One example is the human consumption of horse meat (illegal in California) and the ban in France on dwarf tossing, in spite of the opposition by dwarfs who were paid for being tossed. Other examples include profiteering after disasters, selling pollutions permits, and the commercialization of human organs. In these examples, psychological barriers to efficient policies seem as strong as technological ones.

The importance of psychological constraints produced by an ability to rationalize are important to practitioners. For example, Statement of Accounting Standards No. 99 (SAS99) identifies the ability to rationalize fraud as a central risk factor: “[t]hose involved in a fraud are able to rationalize a fraudulent act as being consistent with their personal code of ethics.”² To appreciate the economic relevance of rationalization one need only consider the social cost of fraud. Buckoff (2001) reports that employee

¹David Brooks (2008) writes in the New York Times: “In reality, we voters — all of us — make emotional, intuitive decisions about who we prefer, and then come up with post-hoc rationalizations to explain the choices that were already made beneath conscious awareness.”

²<http://www.aicpa.org/download/members/div/auditstd/AU-00316.PDF>.

fraud alone costs employers \$400 billion or about 6% of organization's revenues.

In this paper we aim to better understand the underlying logic of rationalization by developing a formal model. Our main premise is that agents choose according to their preferences as in standard economic models, but face (potentially unobservable) psychological constraints in addition to observable incentive and feasibility constraints. So, for example, a manager may have the opportunity and incentive to commit fraud but, absent a convincing rationale that legitimizes such behavior, will choose not to do so.

Informally, in our model, a decision maker (Dee) has preferences over alternatives but, unlike the standard theory, Dee also has a set of *rationales* (modeled as a set of binary relations). Dee chooses the alternative she prefers from among the feasible options she can *rationalize* i.e., those that are optimal according to at least one of her rationales. The ability to rationalize a choice is, therefore, the ability to find a rationale which can justify that choice (because, according to that rationale, no other option is better). Hence, like the standard theory of choice, rationalization is a constrained optimization process.

Consider the following scenario. Dee decides to take time off from work to see a movie. However, prior to leaving the office she is informed that a colleague is in the local hospital and can accept visitors that afternoon. Dee reconsiders her decision to go to the movie and, instead, stays at work.

Suppose that x denotes attending the movie, y denotes staying at work and z denotes visiting the colleague at the hospital. Dee's behavior can be produced by our model of rationalization as follows. Dee prefers x to y to z (i.e., she prefers the movie to work and work to visiting the hospital). Dee has two rationales available to justify her choices. Under rationale 1 Dee's work is pressing so y is ranked above x and z . Under rationale 2 work is not pressing but Dee must visit the hospital rather than go to the movie. So, rationale 2 ranks z above x and x above y . In a binary choice Dee chooses the movie over work because she prefers the movie to work and can rationalize this choice using rationale 2. However, if Dee must choose between all three options she chooses work because she can't rationalize her preferred choice (the movie) over the hospital but she can rationalize her second choice of work (by rationale 1).

Dee's behavior in the example cannot be accommodated by standard economic theory. As is well-known, observed choice is consistent with the standard theory if

and only if *the Weak Axiom of Revealed Preferences (WARP)* holds (see Samuelson (1938a) and (1938b)).³ A variety of behavioral *anomalies* (i.e., violations of WARP) have been documented in field and laboratory experiments. These anomalies are not mere curiosities, but seem central to understanding behavior in areas such as contribution to public goods, voting, and marketing. Often these anomalies are taken as evidence that individuals don't have well-ordered preferences. In contrast, rationalization theory can accommodate a wide variety of behavioral anomalies. Hence, these anomalies are consistent with the basic principle in economics of (constrained) optimization of a single stable, well-defined preference relation. Indeed, Dee's preference relation can be a regular order even if the observed choices are cyclic. We now provide a brief informal description of some potential applications of rationalization theory.

1.1. Motives

There is a great deal of practical interest in determining motivations. Actions are treated differently, e.g., in court, depending upon perceived motivations. The motivations of historical figures are debated long after their death. Psychologists provide post-hoc explanations of motivations based on the context of the experimental evidence.⁴ Motivations, however, are not directly observable. Inferences of motivations require a model of behavior. In standard economic theory, motives are captured by the formalism of a preference relation and choice is, essentially, equated with preference. Outside of economics the relationship between choice and motivation is more subtle.

Consider the study of Snyder et al. (1979). They design an experiment intended to demonstrate that subjects have an aversion to handicapped people. They observe that subjects demonstrate no preference between two movies. However, when given the choice of watching a movie alone or in the company of a handicapped person, subjects will choose to watch the movie with the handicapped person when the two movies are the same but watch the movie alone when the two movies are different. The authors conclude that subjects have an aversion to handicapped people but sometimes

³WARP states that if x is chosen over y then y is not chosen, in another issue, if x is available. We use the acronym WARP to replace "the Weak Axiom of Revealed Preference."

⁴These explanations may include an informal version of rationalization. For example, Shafir, Simonson, and Tversky (1993) argue that "decision makers often seek and construct reasons in order to resolve the conflict and justify their choice, to themselves and to others."

feel constrained not to act on that aversion when their action would overtly imply the aversion.

To illustrate how rationalization theory can produce the behavior in the handicap aversion study suppose that, given the choice between watching movie 1 alone (option x) and watching movie 2 alone (y), Dee chooses x . When given the choice between watching movie 1 alone and watching it with someone in a wheelchair (option z) Dee chooses z . Finally, given the choice between y (watching movie 2 alone) and z (watching movie 1 with a person in a wheelchair), Dee chooses y . Dee's observed choices form a cycle: x is chosen over y , y is chosen over z and z is chosen over x .

Under the handicap aversion thesis Dee prefers x to z even though she chooses z rather than x . Rationalization theory can formally capture both Dee's behavior and handicapped aversion as follows. Dee prefers x to y to z that is, she prefers to avoid the handicapped. However, while Dee can rationalize the choice of x over y and y over z she cannot rationalize the choice of z over x . That is, she cannot rationalize watching movie 1 alone rather than watching movie one with the person in a wheelchair. Even though Dee's observed choices are cyclic it may be that her preferences are not.

The example above demonstrates that an inability to rationalize handicapped aversion can explain Dee's behavior. However, there are other theories that might explain the observed behavior equally well. For example, perhaps Dee's preferences are cyclic and she can rationalize any choice.

Given that psychological constraints are entirely unobservable inferring handicapped aversion specifically and motivations more broadly is problematic. Indeed, our first result demonstrates is that if psychological constraints are unobservable and no restrictions are put on the possible form of such constraints then unique inference of motivation is impossible. Without placing further restrictions on psychological constraints it is simply impossible to tell whether a choice of x over y reflects a preference for x or a psychological constraint against choosing y .

One contribution of rationalization theory is that it implies additional logical restrictions on psychological constraints. We show that rationalization theory implies that if an alternative is psychologically feasible in a superset then it must also be feasible in any subset that contains it. Moreover, any specification of psychological constraints that satisfies this property can be generated as the consequence of some appropriately selected set of rationales that are asymmetric and transitive. As

a result, rationalization theory allows conclusive inferences about preferences even when psychological constraints are entirely unobservable. Basic rationalization theory allows a conclusive inference of preference follows *only* from a violation of WARP. It is precisely when standard theory leads to a contradictory inference of preferences that rationalization theory leads to a conclusive inference of preference. Further inferences are possible if preferences are assumed to be orders. In that case it is sometimes possible to uniquely determine a preference order even in the presence of multiple violations of WARP. In the appendix we provide a list of observed anomalies which may be consistent with rationalization theory.

Returning to the example of Synder’s discrimination study, we show that rationalization theory alone is insufficient to permit the inference of handicapped aversion. We show that the inference of handicapped aversion requires not only the assumption that decision makers have preference orders but also some specific ad hoc assumptions about what choices Dee can rationalize. Specifically, it requires the ad-hoc (but natural) assumption that Dee can rationalize watching movie 1 with a person in a wheelchair rather than watching movie 2 alone. Hence, even partial knowledge of rationales can provide additional insight into preferences.

We fully characterize the inferences of preferences that must follow from rationalization theory aided by ad-hoc provisos on psychological constraints. In a discussion section following the presentation of formal results we discuss potential methods for ascertaining information about agent’s rationales.

Finally, given that rationalization theory and standard theory reveal preferences under different behavioral patterns, we suggest a synthesis that combines both theories to broaden the scope of choice functions amenable to complete identification of preferences. We achieve this synthesis by focusing on sets of rationales that impose the minimal constraints required to accommodate observed choices. That is, we use psychological constraints as parsimoniously as possible. Under the *minimum constraint principle*, when choice is not anomalous, revealed preferences under rationalization theory coincide with preferences revealed by standard theory. When observed choices are acyclic, but possibly anomalous behavior, we show that Dee’s preference order is completely identified by her binary choices. If Dee chooses x over y in a binary choice then she must prefer x over y . Hence, the domain of behavioral patterns amenable to complete identification of preferences is indeed expanded. However, complete identification of preferences is not obtained when observed choices are cyclic. In particular,

it does not follow that Dee must prefer to avoid the handicapped given her observed cyclic choices.

The organization of the paper is as follows. In section 2, we provide a brief review of the related literature. Section 3 introduces the basic formalization of rationalization theory including psychological constraints and what it means to infer preferences. In section 4, we characterize the scope of rationalization theory. Section 5 consists of a discussion of obtaining data about rationales and the implications of rationalization theory for behavior modification and political debate. Section 6 provides a summary conclusion. All proofs are in the appendix which also contains a discussion of application of rationalization theory to explain a variety of behavioral anomalies and ideas about how to incorporate rationalization theory into game theory.

2. Related Literature

A growing literature focuses on accommodating anomalies as a product of internal conflicts. Kalai, Rubinstein and Spiegler (2002) consider a basic model of multiple selves, where choice is optimal according to one of the selves. Green and Hojman (2007) develop a multiple-self model that has no empirical content, but allows partial inferences of preferences. A literature review on multiple-self models can be found in Ambrus and Rozen (2008) who also develop a multiple-self model.

In our approach, Dee has several rationales, but only one preference relation. So, it is straightforward to incorporate rationalization into economic analysis because payoffs need not reflect multiple objectives. Moreover, welfare analysis is transparent because Dee has a unique objective function.

We show that the basic model of rationalization theory implies choice behavior that is consistent with weakening of WARP. Manzini and Mariotti (2007a, 2009) develop a theory that is consistent with the same observable behavior and coin the term Weak WARP. Hence, the scope of rationalization theory is related to the scope of their alternative theory.

The word “rationalizability” is used in game theory (see Bernheim (1984), Pearce (1984)).⁵ Informally, a strategy can be rationalized if some belief about the play of others make that strategy a best-response. Our model and the Bernheim and Pearce notion of rationalizability are very different but share a common idea that actions

⁵See also Sprumont (2000) for another use of the word “rationalizable” in game theory.

can be taken when justified by an argument.

Psychologists who study cognitive dissonance also use the word “rationalization” differently from us. Their basic claim is that people devalue rejected choices and upgrade chosen ones (see Chen (2008) for a review). In the area of motivated cognition, Von Hippel (2005) provides a survey on self-serving biased information processing (see also Akerlof and Dickens (1982), Rabin (1995), Dahl and Ransom (1999), Carrillo and Mariotti (2000) and Bénabou and Tirole (2002)).

A large literature deals informally with rationalization in political science. For example, Achen and Bartels (2006) argue that voters justify their support for candidates by discounting unfavorable data. Mendelberg (2001) claims that policy arguments effectively allow voters to rationalize racial prejudice.

An early literature review on psychology and economics can be found in Simon (1959). More recently, Rabin (1998) and Mullainathan and Thaler (2000) provide surveys on models that explain violations of expected utility theory.⁶

3. Basic Concepts

Let A be a finite set of alternatives. A non-empty subset $B \subseteq A$ of alternatives is called an *issue*. Let \mathcal{B} be the set of all issues. A *choice function* is a mapping $C : \mathcal{B} \rightarrow A$ such that $C(B) \in B$ for every $B \in \mathcal{B}$. Hence, a choice function takes an issue as input and returns a feasible alternative (i.e., the choice) as output.

A binary relation R on A is called a *rationale*. A *preference* is an asymmetric rationale. A transitive, complete preference is an *order*. By standard convention, $x R y$ denotes that, x is R -preferred to y . A decision maker, named Dee, makes choices C . Another agent, named Bob, observes Dee’s choices and make inferences about Dee’s preferences.

Standard economic theory equates revealed preferences and observed choice. If choices and revealed preferences are not equated by definition then the question of how preferences are inferred from choice becomes open to new conceptual and empirical analysis. In particular, when an observer reports his inferences of preferences from choice it is possible to test these inferences against a theory of inferred preferences. For example, if Bob claims that Dee’s choice of x when y is available does not reveal

⁶See also Gul and Pesendorfer (2005) for a critique of neuroeconomics that raises broader concerns about integrating psychological models of decision-making into economics.

a preference for x then we would like to be able to determine whether Bob's claim is consistent with rationalization theory.

Our focus is on the development of rationalization theory and on the preference inferences that this theory allows. As mentioned in the introduction, the core idea of rationalization is that there are psychological constraints on choice. We formalize such constraints by defining a *psychological constraint function* : a mapping $\tau : \mathcal{B} \rightarrow \mathcal{B}$ such that $\emptyset \neq \tau(B) \subseteq B$ for every issue $B \in \mathcal{B}$. A psychological constraint function takes an issue as input and returns, as output, the set of all alternatives in the issue that are psychologically feasible. Dee chooses the option that she prefers among all alternatives that are psychologically feasible. Formally, a *model of behavior* is a pair (R, τ) of a preference and a psychological constraint function. A model of behavior (R, τ) *underlies* a choice function C if for any issue $B \in \mathcal{B}$, $C(B) \in \tau(B)$ and

$$C(B) R y \text{ for all } y \in \tau(B), y \neq C(B).$$

So, Dee chooses as if she solves a constrained optimization problem with psychological constraints. The standard economic model is the special case without psychological constraints, i.e., $\tau(B) = B$.

Psychological constraints may produce a formal model in which Dee does not choose the alternative she most prefers for a given issue. Any pattern of observed choice can be accommodated by an appropriately chosen constraint function. This follows because any choice of x in issue B can be accommodated by a model in which Dee's choice is dictated by Dee's constraint, i.e., $\tau(B) = \{x\}$. As a consequence any theory that allows preference inferences must put restrictions on psychological constraints. Below we show that rationalization theory imposes constraints and argue that observers typically, albeit implicitly, make ad hoc assumptions about psychological constraints.

Let Υ be the set of all psychological constraint functions. Let \mathcal{P} be the set of all preferences. A *theory of behavior* is a subset $\hat{\mathcal{P}} \times \hat{\Upsilon} \subseteq \mathcal{P} \times \Upsilon$ of psychological constraint functions and preferences. So, in general, a theory of behavior is a collection of models of choice, defined by provisos on preferences and/or psychological constraints.

We first explore a qualification on psychological constraints that follows from the concept of rationalization. Intuitively, an alternative $x \in B$ can be rationalized only if Dee can find a compelling way (in her judgement) to tell herself that x is an acceptable

choice. To determine what is acceptable, Dee must have ways, perhaps multiple ways, to compare alternatives. This can be formalized by the concept of rationales. Given a rationale R_i , $x R_i y$ denotes that, by R_i , x is acceptable in the presence of y . Given an issue B , an option $x \in B$ is *rationalized* by R_i if $x R_i y$ for all $y \in B$, $y \neq x$. That is, no physically feasible alternative y makes x unacceptable according to rationale R_i . Let $\mathcal{R} = \{R_i, i = 1, \dots, n\}$ be the set of all rationales that Dee finds compelling. Given \mathcal{R} , an option $x \in B$ is *rationalizable* in B if x can be rationalized by some rationale $R_i \in \mathcal{R}$. A set of rationales \mathcal{R} determines the psychological constraint function $\tau^{\mathcal{R}}$, where $\tau^{\mathcal{R}}(B)$ are the rationalizable options in B .

Preliminary result For any set \mathcal{R} of rationales, the psychological constraint function $\tau^{\mathcal{R}}$ satisfies

$$\text{if } B \subseteq B^* \text{ then } \tau(B^*) \cap B \subseteq \tau(B) \quad (3.1)$$

Moreover, if a psychological constraint function τ satisfies 3.1 then there exists a set \mathcal{R} of rationales (where each rationale in \mathcal{R} can also be shown to be transitive and asymmetric) such that $\tau = \tau^{\mathcal{R}}$.

The preliminary result shows that rationalization can be understood as requiring that psychological constraints satisfy 3.1. Let's say that a pair of issues $(B, B^*) \in \mathcal{B} \times \mathcal{B}$ is *nested* if $B \subseteq B^*$; B is the *sub-issue* and B^* is the *super-issue*. If option x is psychologically feasible in the super-issue B^* then no alternative in B^* (and, hence, no alternative in B) makes x psychologically unacceptable. So, x must also be psychologically feasible in the sub-issue B . Informally, rationalization theory implies that if Dee can rationalize choosing x from some larger set of alternatives then she can also rationalize choosing x from any subset of those alternatives.

Let $\Upsilon^1 \subseteq \Upsilon$ be the set of psychological constraint functions that satisfy 3.1. These are the psychological constraints coming from the inability to rationalize.⁷ Let $\mathcal{P}_x \Upsilon^1$ be the basic *theory of rationalization*.

Let $\mathcal{P}^o \subseteq \mathcal{P}$ be the set of preferences orders. We define $\mathcal{P}^o_x \Upsilon^1$ as *theory of order rationalization*, i.e., rationalization theory with the additional restriction that preferences are orders. We also consider further restrictions on psychological constraints.

⁷The preliminary result also shows that requiring transitivity and asymmetry on rationales does not lead to additional qualification on psychological constraint functions. However, if rationales must be orders then the psychological constraint functions are more restricted than Υ^1 .

The motivation here is that Bob’s inferences about Dee’s preferences may depend critically on ad-hoc assumptions that he makes about what is psychologically feasible for Dee. These assumptions may be based on subtle factors that economists tend to find of secondary (if any) importance, but that have been the focus of great interest elsewhere across the social sciences. These factors include what Bob knows about Dee’s cultural background and also on what Dee says to Bob. For example, Bob may assume that Dee does not face any psychological constraints when choosing between “decent” movies, but what constitutes decent themes may depend, in part, on Dee’s cultural background. In addition, if Dee tells Bob that she finds eating chocolate (z) healthier than eating peanuts (y) then Bob may assume that Dee can rationalize z over y . After all, Dee is essentially providing Bob with a rationale for choosing chocolate over peanuts and so, Bob may assume that for Dee z is psychologically feasible in $\{y, z\}$.

Formally, let $\mathcal{A} = \{(y_i, B_i); y_i \in B_i \ i = 1, \dots, n\}$ be a collection of n issues $B_i \in \mathcal{B}$ and alternatives $y_i \in B_i$. We interpret \mathcal{A} as ad-hoc assumptions that $y_i \in B_i$ is psychologically feasible in issue B_i , $i = 1, \dots, n$. Let $\Upsilon^{\mathcal{A}} \subseteq \Upsilon^1$ be the set of all psychological constraint functions τ that satisfy 3.1 and such that $y_i \in \tau(B_i)$, $i = 1, \dots, n$. In these psychological constraint functions of rationalization theory, z_i is psychologically feasible in B_i . Let $\mathcal{P}_x \Upsilon^{\mathcal{A}}$ be the *theory of \mathcal{A} -rationalization* and $\mathcal{P}^o_x \Upsilon^{\mathcal{A}}$ be the *theory of order \mathcal{A} -rationalization*.

4. Main Results

A choice function C is *consistent* with a theory of behavior $\hat{\mathcal{P}}_x \hat{\Upsilon}$ if there exists a model of behavior $(R, \tau) \in \hat{\mathcal{P}}_x \hat{\Upsilon}$ that underlies C . So, a choice function is consistent with a theory if the observed choices can be accommodated by a model of the theory.

Definition 1. Given choice function C and theory $\hat{\mathcal{P}}_x \hat{\Upsilon}$, Bob *infers* that Dee prefers x to y if C is consistent with $\hat{\mathcal{P}}_x \hat{\Upsilon}$ and x is preferred to y ($x R y$) in every model of behavior $(R, \tau) \in \hat{\mathcal{P}}_x \hat{\Upsilon}$ that underlies C .

So, Bob infers that Dee prefers x over y if x ranks higher than y in every model of behavior (within the paradigm permitted by Bob’s theory) that underlies Dee’s choices. Let us now consider the preference inferences that must follow from the basic rationalization theory. Recall that a pair of nested issues $(B, B^*) \in \mathcal{B} \times \mathcal{B}$

violates WARP if

$$B \subseteq B^*, C(B^*) \in B, \text{ and } C(B) \neq C(B^*).$$

So, in the super-issue B^* , $C(B^*)$ is chosen over $C(B)$ and in the sub-issue B , $C(B)$ is chosen over $C(B^*)$. By definition, these choices are an anomaly. To simplify the language, let's say that (B, B^*) is *anomalous* if it is a pair of nested issues that violate WARP. In an anomaly, Dee must prefer $C(B)$ to $C(B^*)$ because in the sub-issue B , Dee chooses $C(B)$ over $C(B^*)$ and she must be able to rationalize $C(B^*)$ in B because in the super-issue B^* she chose $C(B^*)$ when *all* alternatives in B were available. Formally,

Proposition 1. *Consider a choice function C consistent with the basic rationalization theory $\mathcal{P}\chi\Upsilon^1$. Bob infers that Dee prefers x over $y \neq x$ if and only if there is an anomaly (B, B^*) such that $x = C(B)$ and $y = C(B^*)$.*

In standard theory, an anomaly leads to contradictory inferences over preferences. In contrast, Proposition 1 shows that through anomalies, and only through anomalies, rationalization theory delivers *conclusive* and non-contradictory inferences over preferences. As mentioned above this follows because, in rationalization theory, a choice by itself may be the consequence of a preference or a psychological constraint. Under rationalization, a choice of y in the super-issue reveals the absence of a constraint against y in the sub-issue and the sub-issue. Hence, the choice of x in the sub-issue a preference for x over y . In addition, proposition 1 reveals exactly where the link between choice and preference is broken. In the super-issue of any anomaly, Dee makes a choice $C(B^*)$ even though there exists a feasible alternative that she prefers to $C(B^*)$.

Proposition 1 shows whether the psychological explanations provided for observed choices are examples of rationalization models of behavior that accommodate the observed choices or are logical conclusions that must follow from rationalization theory. In the contribution to public goods study of Berger and Smith (1997), a small contribution (x) preferred to no contribution (y), necessarily follows from rationalization theory, but no contribution (y) preferred to a large contribution (z) does not.

Recall the discrimination study of Snyder et al. (1979). There are three alternatives: to watch movie 1 alone (x); to watch movie 2 alone (y); and to watch movie

1 with a handicapped person in a wheelchair (z). Let's assume that, as observed by Snyder et al., some subjects watch movie 1 with the handicapped person over watching movie 1 alone (i.e., $\bar{C}(x, z) = z$). The central claim in Snyder et al. is that some of these subjects prefer to avoid the handicapped. That is, they prefer x to z , even though they chose z to x . This claim is seemingly backed if the alternative movie (2) is always chosen when movie 1 with the handicapped person is an option (i.e., $\bar{C}(y, z) = y$ and $\bar{C}(x, y, z) = y$) even though movie 1 is chosen over movie 2 (i.e., $\bar{C}(x, y) = x$). Snyder et al. interpretation is that subjects prefer to avoid the handicapped, but can only rationalize this choice when the movies are different. However, proposition 1 shows that the only preference inference that necessarily follows from the basic theory of rationalization is that x is preferred to y , i.e., the uninteresting conclusion that movie 1 is preferred to movie 2. Now consider the additional assumption that Dee can rationalize watching the movie with the handicapped (e.g., she can rationalize z over y). This assumption seems completely natural given our current cultural background. Then, the choice of y over z now leads to the conclusion that Dee prefers y to z . Finally, if Dee preferences form an order then Dee must prefer to avoid the handicapped, i.e., she prefers x to z .

Claim 1. *Given a cyclic choice function \bar{C} , Bob infers that Dee must prefer watching the movie alone to watching the movie with the handicapped person (i.e., xRz) under the basic rationalization theory and two additional assumptions*

1. Dee's preferences form an order;
2. Dee can rationalize watching the movie with the handicapped.

Claim 1 describes the formal assumptions necessary to uniquely infer handicapped avoidance on the basis of the behavior observed in the Snyder et. al., study. *An important implication here is that the ability to discriminate depends crucially upon the availability of rationale that legitimizes the discriminatory choice.* It is possible that discriminatory actions are not taken because of psychological constraints, but will be taken once again if the set of accepted rationales are expanded to include a legitimization of discrimination. In the opposite direction, discriminatory actions maybe reduced as it becomes de-legitimized even if preferences for discrimination remain unchanged. This follows whenever de-legitimization creates a psychological barrier to discrimination.

In addition to illustrating how behavior may change in the absence of any changes in preferences, the Snyder et al. study delivers a simple way to show non-trivial inferences that comes from speech under rationalization theory. Assume that Bob adopts order rationalization theory, but does not assume that Dee can rationalize watching the movie with the handicapped. By proposition 1, given the choice function \bar{C} , Bob cannot infer that Dee prefers to avoid the handicapped. However, if Dee reveals to Bob that she faces no psychological constraint in watching a movie with someone in a wheelchair. Then, by Claim 1, Bob must infer that Dee prefers to avoid the handicapped.

Finally, the Snyder et al. study also gives us a good opportunity to show how rationalization theory can aid in the interpretation of experimental data. The choice function \bar{C} form a cycle which, by standard theory, cannot be reconciled with preference orders. However, it is difficult to appreciate without formal theory that the assumption of preference orders is required for the conclusion reached by Snyder et al. of a hidden preference for avoiding the handicapped. In the next sections, we show that assumptions 1 and 2 (of orders and ad-hoc provisos) are not only sufficient in Claim 1, but they are also necessary.

4.1. The need for orders

Given a choice function C , let \tilde{R}^C be a binary relation defined by the binary choices. That is, $x \tilde{R}^C y$ if and only if $C(\{x, y\}) = x$.

Proposition 2. *Consider a model $(R, \tau) \in \mathcal{P}_x \Upsilon^1$ that underlies a choice function C . Then, the model (\tilde{R}^C, τ) also underlies the choice function C .*

By proposition 2, if a choice function is consistent with rationalization theory then it is always possible to accommodate Dee's choices by preferences \tilde{R}^C determined by her binary choices. This leads to the following corollary.

Corollary 1. *Consider a choice function C such that x is chosen over y , i.e., $C(\{x, y\}) = x$ and any $\mathcal{P}_x \Upsilon^A$ theory of \mathcal{A} -rationalization. Then, Bob does not infer that Dee prefers y to x .*

So, if Dee chooses x over y in a binary choice and Dee's preferences are *not* assumed to be orders then it does not follow that Dee must prefer y to x no matter

which ad-hoc psychological assumptions are made over what Dee can be rationalize. In particular, the demonstration that Dee prefers to avoid the handicapped *requires* the assumption of preference orders and cannot be obtained by *any* ad-hoc provisions alone.

4.2. Minimum constraint principle and ad-hoc assumptions

It is easy to check that Bob does not infer that Dee prefers to avoid the handicapped if the ad-hoc assumption (2) is disregarded in Claim 1. However, one may suspect that this follows because the basic rationalization theory is parsimonious about preference revelation. For example, assume that Dee chooses beef over ham. Standard theory concludes that Dee prefers beef over ham, but the basic rationalization theory does not. It allows for the possibility that Dee prefers ham and her choice is, perhaps, a reflection of a religious restriction that does not allow the rationalization of ham. In particular, there are choice functions, notably those that satisfy WARP, for which rationalization theory does not make any inferences over preferences. In this section, we propose a fusion of rationalization theory and the standard theory of choice which will allow for a wider variety preference inferences. Even still, we will be unable to conclude that Dee prefers to avoid the handicapped in the cycle of \bar{C} .

Let's start with a basic philosophical principle in economics: the idea that, in the absence of additional information, Dee's choice is unconstrained and, hence, her choice reveals her preference.⁸ We adopt this idea here, but only to the extent that it does not contradict Dee's observed choices. That is, we use an "Occam razor" principle of minimum deviations from standard economic analysis that still allow us to accommodate the observed choices. This idea can be formalized by psychological constraint functions that satisfy 3.1; are as unrestrictive as possible; and can underlie the choices.

Fix a choice function C that is consistent with rationalization theory $\mathcal{P}\mathbf{x}\Upsilon^1$. Let (R, τ) and (R', τ') be two models that underlie the choice function C . The model (R, τ) is *dominated* by the model (R', τ') if R' is an order, and $\tau(B) \subseteq \tau'(B)$ for all issues $B \in \mathcal{B}$, with strict inclusion for some issue $B \in \mathcal{B}$. So, if a model (R, τ) is

⁸The ideas in this principle can be traced back to the libertarian school of thought that opposes paternalistic policies based on welfare criteria that are inconsistent with the notions of well-being upheld by the individuals affected by these policies (see, for example, Mills (1860)). See also Thaler and Sunstein (2003) for a discussion of paternalism and modern behavioral economics.

dominated then it uses more constraints than necessary to accommodate the observed choices.

Given a choice function C , let $\mathcal{P}^C_{\mathbf{x}\Upsilon^C}$ be the *minimum constraint theory of rationalization*: the set of all models $(R, \tau) \in \mathcal{P}_{\mathbf{x}\Upsilon^1}$ that underlie C and are not dominated. These models do not require more constraints on choice than needed to underlie Dee's choices.

As noted above, the minimum constraint theory of rationalization is based on a criteria that is implicit in standard economics. However, the minimum constraint principle can also be motivated by the common perception in psychology that people can easily rationalize and only on occasion find themselves unable to do so. We now show inferences over preferences made under minimum constraint rationalization theory.

We say a choice function is *cyclic* if there are three alternatives (say x , y and z) such that,

$$C(\{x, y\}) = x; \quad C(\{y, z\}) = y; \quad \text{and} \quad C(\{x, z\}) = z.$$

A choice function that is not cyclic is *acyclic*.

Proposition 3. *Let C be an acyclic choice function that is consistent with rationalization theory. If $C(\{x, y\}) = x$, $x \neq y$ (i.e., x is chosen over y) then Bob infers that Dee prefers x to y by minimum constraint rationalization theory.*

Proposition 3 shows that when a choice function is acyclic, binary choices are revealed as the unique preference order implied by minimum constraint rationalization theory. Among all possible preferences (orders or not), the only surviving preference is the order given by the binary choices. *So, minimum constraint rationalization theory strictly extends the domain of choice functions from which preferences can be completely identified.* Indeed, if the choice function satisfies WARP then it is acyclic and so, the revealed preference is the same as in standard theory. However, complete identification of preferences now extends to cases where choices may be anomalous.

Consider the following basic question: when can Bob infer that x is preferred over y after observing a binary choice of x over y ? Proposition 3 delivers a simple and compelling answer: *Under minimum constraint rationalization theory, if no cycles are observed then, whether or not Dee's behavior is anomalous, Dee's binary choices reveal her preferences.*

However, when binary choices are cyclic, Bob cannot infer a complete preference relation by minimum constraint rationalization theory. It is easy to check that more than one undominated model of behavior can accommodate a cycle. Hence, the basic economic principle of assuming choice to be as unconstrained as possible is insufficient to reveal all preferences in the cycle. In addition, in the cycle of \bar{C} , *it does not follow that Dee prefers to avoid the handicapped by minimum constraint rationalization theory, even if Dee is assumed to have preference orders.* We now show that ad-hoc provisos are needed for this demonstration.

4.3. Revealed preferences

Given the significance of (order) \mathcal{A} -rationalization theory it is natural to ask what preferences are revealed by orders and/or additional psychological assumptions. Propositions 3 and 4 deliver a full characterization of what Bob can infer about Dee's preferences by (order) \mathcal{A} -rationalization theory.

Proposition 4. *Consider a choice function C consistent with \mathcal{A} -rationalization theory $\mathcal{P}_x\Upsilon^{\mathcal{A}}$. Then, Bob infers that Dee prefers x to y , $x \neq y$, if and only if at least one of the two conditions hold : 1) Bob infers that Dee prefers x to y by basic rationalization theory or 2) $x = C(B)$ and for some $(y_i, B_i) \in \mathcal{A}$, $y = y_i \in B \subseteq B_i$.*

Take the basic rationalization theory as a benchmark. Proposition 3 shows that the *only* additional inferences over preference that follows from ad-hoc assumptions are the natural ones: if Bob assumes that Dee can rationalize y_i in B_i then he must infer that Dee prefers her choice $C(B)$ over y_i .

Proposition 5. *Consider a choice function C consistent with order \mathcal{A} -rationalization theory $\mathcal{P}^o_x\Upsilon^{\mathcal{A}}$. Then, Bob infers that Dee prefers z_1 to z_k if there exists a chain z_{i+1} , $i = 0, \dots, k - 1$, such that Bob infers that Dee prefers z_i to z_{i+1} by \mathcal{A} -rationalization theory.*

Take \mathcal{A} -rationalization theory as a benchmark. Proposition 5 shows that the *only* additional inferences that come from orders are also the natural ones: if Bob concludes that Dee prefers x to y and y to z then he must also conclude that Dee prefers x to z .

5. Testing rationalization theory

In this section, we characterize the choice functions that are consistent with rationalization theory. Given a choice function C and a set \mathcal{A} of psychological assumptions, let $R^{C,\mathcal{A}}$ be the binary relation defined by $x R^{C,\mathcal{A}} y$ if and only if

$$\begin{aligned} x &= C(B) \text{ and } y = C(B^*) \text{ for a pair } (B, B^*) \text{ of nested issues that violate WARP or} \\ x &= C(B) \text{ and for some } (y_i, B_i) \in \mathcal{A}, y = y_i \in B \subseteq B_i. \end{aligned}$$

By proposition 4, $R^{C,\mathcal{A}}$ is the binary relation of preferences revealed by \mathcal{A} -rationalization theory. In the appendix, we show that a choice function C is consistent with \mathcal{A} -rationalization theory if and only if $R^{C,\mathcal{A}}$ is asymmetric. Moreover, a choice function C is consistent with order \mathcal{A} -rationalization theory if and only if $R^{C,\mathcal{A}}$ is acyclic. These results fully characterize the empirical content of (order) \mathcal{A} -rationalization theory. We now reformulate these results in a standard axiomatic form for the basic rationalization theory (i.e., $\mathcal{A} = \emptyset$ and preferences are not necessary orders).

Consider two pair of nested issues (B_1, B_1^*) and (B_2, B_2^*) that violate WARP. The choices on these two nested issues are *reversed* if $C(B_1) = C(B_2^*)$ and $C(B_1^*) = C(B_2)$. So, the choice in the sub-issue of one of the pairs is the choice on the super-issue of the other pair. By proposition 1, a reverse pair of nested issues would imply, by rationalization theory, that $C(B_1)$ is revealed preferred to $C(B_1^*)$ and that $C(B_1^*)$ is revealed preferred to $C(B_1)$. So, $R^{C,\emptyset}$ is *not* asymmetric.

Irreversibility A choice function C satisfies the irreversibility axiom if there are no two pairs of nested issues that violate WARP with reversed choices.

So, the irreversibility axiom rules out contradictory inferences by rationalization theory. This axiom fully demarcates the choice functions that can and cannot be accommodated by rationalization theory. The irreversibility axiom can also be restated in familiar terms.

WWARP A choice function C satisfies the weak weak axiom of revealed preferences iff

$$x \neq y, \{x, y\} \subseteq B_1 \subseteq B_2, C(\{x, y\}) = C(B_2) = x \text{ then } C(B_1) \neq y.$$

The WWARP states that if x is chosen over y (in the binary choice) and x is also chosen over y (in an issue B) then y is not chosen over x in a sub-issue of B .

Proposition 6. *The irreversibility axiom holds if and only if WWARP holds.*

The equivalence between irreversibility and WWARP leads to corollary 1.

Corollary 2. *A choice function C is consistent with rationalization theory if and only if it satisfies WWARP.*

The empirical content of standard theory of choice and rationalization theory are each based on a single axiom (WARP and WWARP) and these two axioms are directly related to each other: WWARP is a familiar and natural relaxation of WARP (see Manzini and Mariotti (2007a)). So, as we noted, rationalization theory can accommodate a wide variety of anomalies. Hence, violations of WARP do not imply a violation of the basic principle in economics that Dee behaves as if she follows a constrained optimization process of a single stable preference. In addition, for these anomalies, this preference can be an order.

5.1. Interpreting (Perhaps Contradictory) Speech

Assume that Dee tells Bob that, if given the opportunity she should (and would) choose x over y . Bob may interpret this speech in different ways. The two polar cases are (1) to believe Dee's speech and (2) ignore this data. Both approaches are unsatisfactory. On the one hand, Dee may be lying in the sense that when facing the actual choice between x and y Dee may choose y . On the other hand, there may be relevant information in Dee's statement. Rationalization theory offers a third choice. Bob may interpret Dee's statement as a statement about rationales. If Dee says that she should choose x over y then Bob may interpret this statement as revealing that Dee can rationalize the choice of x over y . It does not mean that Dee will choose x over y only that if Dee chooses y then she necessarily prefers it to x . The substantive point here is that rationales may be inferred not only from observed choice, but also from observed speech. So, rationalization theory can be used as a formal way to interpret speech. In addition, Bob can leverage very limited knowledge of rationales available to Dee to help infer preferences from observed choice. Hence, agents' explanations for their behavior may become informative about their underlying preferences.

An interesting implication of this discussion is that contradictory speech may also be informative. Assume that Dee indicates that x could be chosen over y and also that y could be chosen over x . These apparently contradictory statements may be interpreted to mean that Dee can rationalize both options. To return to our example in the introduction, consider an accountant who says that she could never do anything illegal, but she also says that the company makes too much money and she is not properly compensated by her work. So, if the choices are to engage in fraud (x) or not (y) then her contradictory statements may be interpreted as the ability to rationalize both options.

5.2. Changing Rationales, Behavior and Social Norms

The academic economics literature has emphasized that behavior changes in response to changes in opportunity and incentives. Rationalization theory provides a formal framework to understand an additional mechanism for changes in behavior. Behavior may change as a consequence of a change in acceptable rationales. This mechanism provides an explanation for the effort that leaders often spend on communicating “core values”. Leaders may alter behavior by adding and/or subtracting rationales, without necessarily changing preferences, opportunities or incentives. This insight is also consistent with the idea that while preferences may be unaffected by speech, it is possible that social discourse may change conventional mores (e.g., with respect to marriage, relations between adults and children, and attitudes towards stigmatized behavior). The effect of such speech may be to change people’s rationales and, hence, what is psychologically feasible.

In laboratory experiments, Ariely and Mazar (2006) found that a willingness to cheat depends upon whether subjects thought people like them were cheating. In field experiments, Goldstein and Cialdini (2007) found that hotel guests willingness to reuse towels depended upon whether they were informed that most other guests reuse their towels.

Rationalization theory can help understand how social norms affect behavior and, in particular, why Dee may take costly actions that are difficult to observe. If Dee is informed that most guest reuse their towels then the rationale that “few reuse towels” become less compelling. If she is unable to find a compelling rationale for not reusing her towels then she faces a psychological constraint. Hence, psychological constraints may be a root cause of socially desirable behavior.

Relaxing psychological constraints may also have an impact on behavior. In the 1940s General Mills invented powdered cake mix that allowed homemakers to simply add water and bake. The product saved time, but its poor sales surprised the company. The research of psychologists Burleigh Gardner and Ernest Dichter suggested that powdered eggs should be left out (even though this reduced the convenience of the product) so that fresh eggs could be added by consumers giving them a sense of creative contribution. The solution was successfully implemented and the cake mix gained notoriety. It is difficult to say whether cakes made with fresh eggs simply taste better, but the example shows how psychologists believe that lifting psychological constraints leads to behavioral changes.⁹

6. Conclusion

In this paper we formalize the widely used concept of rationalization. We argue that the essential feature of rationalization is that agents want to take actions they like but are sometimes constrained from doing so because they lack an adequate rationale. The possibility that choice is psychologically constrained makes rationalization economically relevant and appropriate for economic analysis as a process of constrained optimization. Our first contribution is to formally define rationalization and show that rationalization can accommodate several behavioral patterns that are incompatible with the standard theory of choice.

We follow a revealed preference approach and ask, without knowledge of preferences or psychological constraints, what behavior is consistent with constrained optimization. In a general model that imposes no structure on psychological constraints it is impossible to draw any inferences about either constraints or preferences on the basis of choice behavior alone. The basic theory of rationalization imposes structure on psychological constraints that makes it possible to infer unobserved preferences (as well as psychological constraints) from observed choice behavior alone. However, the inferences that can be drawn are relatively limited. Preferences can only be inferred when observed choice behavior violates WARP. Further insight can be obtained if underlying preferences are assumed to be orders. In that case a sequence of violations of WARP can be used to infer preference across a set of decisions in which no violations

⁹Finding Betty Crocker: The Secret Life of America's First Lady of Food, Susan Marks [Simon & Schuster:New York] 2005 (p. 168, 170) <http://www.foodtimeline.org/foodcakes.html>

of WARP are observed.

Standard economic theory infers preference from choice alone. Rationalization theory expands the scope of behavior that may be informative about preference to include speech. We argue that speech may be used to provide data about rationales that help interpret choice behavior in subtle and surprising ways. In the basic rationalization model in which there is no observed violation of WARP it is impossible to determine whether the choice of alternative x over y is the result of a preference for x over y or the consequence of a psychological constraint preventing the selection of the preferred alternative y . The question can be resolved by asking the agent whether someone like him could choose y over x . If the answer is yes then we might reasonably infer the agent is unconstrained in his choice between x and y so that the choice of x reveals a preference.

We use an example to show that in applied work observers often, implicitly, make ad hoc assumptions about the absence of psychological constraints and such assumptions are necessary in order to support inference about preference. Indeed, a polar case of such an ad hoc assumption is implicit in standard choice models: that agents face no psychological constraints on choice. We show that under the assumption of no psychological constraints and preference orders rationalization theory is identical to standard choice theory. Moreover, by slightly weakening the assumption of no psychological constraints to an assumption we call minimum constraint we show that all the inferences one can make with the standard model also follow from rationalization theory. In addition, one can draw inferences about preference even for cases in which WARP is violated.

Rationalization theory can accommodate several behavioral anomalies. Hence, it is natural to ask what behavioral patterns are incompatible with rationalization theory. We provide a full characterization of the empirical content of rationalization theory. We show that the behavioral patterns that can be accommodated by rationalization theory are those that satisfy a natural and familiar relaxation of WARP. Therefore, rationalization theory can be non-parametrically tested in a manner akin to the standard theory.

This paper should be taken as an initial exploration of rationalization. In the introduction we mentioned the application of rationalization theory to accounting practice. Accounting standards explicitly call for auditors to assess the possibility of fraud by considering whether the organization has a culture that might allow ratio-

nalization of fraud. Our theory of rationalization makes explicit just how rationales might impact behavior. It also implies that legitimizing and delegitimizing rationales is an important leadership activity. While our analysis and model is restricted to a decision theory context the rationalization model permits a strategic analysis and suggests why many arguments within organizations and the broader society are debates about rationales and acceptable principles. Debates about the legitimacy of discrimination towards ethnic or religious groups become vital debates with real behavioral implications in the context of the rationalization model whereas in standard economic models such discussions might be viewed as mere cheap-talk. Further exploration of such issues will have to await future work.

7. Appendix

7.1. Behavioral Anomalies

In this subsection we provide a typology of several behavioral anomalies. Given a set of alternatives B , let $C(B)$ be Dee's choice. With three alternatives (say x , y and z), behavioral anomalies can be broken down into three types: *cycles*, *attraction effects* and *difficult choices*.¹⁰

In a *cycle*,

$$C(\{x, y\}) = x; C(\{y, z\}) = y; \text{ and } C(\{x, z\}) = z.$$

In an *attraction effect*,

$$C(\{x, y\}) = x; C(\{y, z\}) = y; C(\{x, z\}) = x \text{ and } C(\{x, y, z\}) = y.$$

In a *difficult choice*,

$$C(\{y, z\}) = y; C(\{x, z\}) = x \text{ and } C(\{x, y, z\}) = z.$$

All three anomalies were repeatedly observed in economically relevant field and laboratory experiments. The nomenclature follows the empirical literature. Cycles

¹⁰Let y be the choice from x , y and z and fix the choice between x and z . So, with three alternatives, there are four distinct behavioral patterns (given by the choices on $\{x, y\}$ and $\{z, y\}$). One of them is consistent with WARP and the other three patterns are distinct anomalies.

have been noted at least since May (1954). Manzini and Mariotti (2007a) provide a review of the empirical literature on cycles. The discrimination study of Snyder, Kelck, Stretna and Mentzer (1979) mentioned above is an example of a cycle.

In the attraction effect an alternative (say z) is not chosen but alters choice. When z is not available Dee chooses x over y , but when z is available Dee chooses y over x and z . Ok et al (2008) provides a survey on the empirical evidence for the attraction effect and also develops a theory that can accommodate it.¹¹ Consider the contribution to public goods study of Berger and Smith (1997). They find that some donors (to universities) make a small solicited contribution (x) over no contribution (y), but if either a small or a large contribution (z) is solicited they do not contribute.

In a difficult choice, an alternative (z) is not chosen in pairwise choices with x and y , but z is chosen when all three choices are available. This anomaly was observed by Tversky and Shafir (1992) in several laboratory experiments (see also Simonson (1989), Simonson and Tversky (1992) and (1993)). In a different field experiment in marketing, Iyengar and Lepper (2000) show that the fraction of customers who bought a gourmet jam was significantly larger when presented with a limited selection than with an extensive selection.

Rationalization theory can accommodate the cycle, the attraction effect and the difficult choice. Informally, rationalization theory accommodates the attraction effect of Berger and Smith (1997) by assuming that Dee prefers to make a small donation over not donating. However, Dee cannot rationalize a small donation when a large donation is also requested, perhaps because then making a small donation no longer seems generous and it cannot be justified on the grounds of maximization of her material benefits either. Informally, rationalization theory accommodates the difficult choice in Iyengar and Lepper (2000) by assuming that Dee prefers any type of jam to not buying any jam. However, Dee cannot rationalize one type of jam over another, unless she is required to buy jam, perhaps because she cannot tell herself why one brand is better than the other.

The attraction effect, the difficult choice and the cycle encompass a large effort by psychologists and economist to empirically demonstrate seemingly odd behavior that violate the standard economy theory of choice. However, these behavioral patterns can be produced by a natural theory of psychological constraints. So, these behavioral anomalies are consistent with the basic economic principle of (constrained)

¹¹See also Masatlioglu and Nakajima (2007), Masatlioglu and Ok (2007), Eliaz and Ok (2006).

optimization of a single preference order.

7.2. Applications to Game Theory

Even in the simple form presented in this paper rationalization theory can be easily incorporated into game theory. Because rationalization theory assumes that Dee has unique preferences, the payoff matrix can be left intact and players optimize given the additional constraints imposed by their rationales. Our objective here is only to demonstrate the kinds of insights such a project might deliver. So, consider the prisoner dilemma with players making rationalized choices. The game is

(I, II)	C	D
C	$(-1, -1)$	$(-20, 0)$
D	$(0, -20)$	$(-10, -10)$

So, 1's preferences are

$$(D, C) \succ (C, C) \succ (D, D) \succ (C, D).$$

and 2's preferences are

$$(C, D) \succ (C, C) \succ (D, D) \succ (D, C).$$

The rationales of player 1 are all orders such that $(C, C) \succ (D, C)$. The rationales of player 2 are all orders such that $(C, C) \succ (C, D)$. That is, none of the players can rationalize defection (D) if the other player cooperates (C). So, if 2 cooperates then 1's feasible options are (C, C) and (D, C) . In this case, 1 cooperates because this is the only rationalizable option. If 2 defects then 1 can rationalize both options (C, D) or (D, D) , and 1 defects because she prefers it. An analogous result holds for player 2. So, the profile in which both players cooperate and the profile in both players defect are now equilibrium outcomes.

Now consider a different game. Assume that player 1 moves first and plays either aggressive (A) or pleasant (P). Player 2 observes the play of 1 and either reciprocates (R) or does not reciprocate (N). Payoffs for (A, R) , (A, N) , (P, R) and (P, N) are $(6, 0)$, $(2, 1)$, $(4, 1)$ and $(3, 2)$, respectively. Player 1's rationales are all orders and players 2's rationales are all orders such that $(P, R) \succ (P, N)$. So, if 1 is pleasant then

player 2 can only rationalize reciprocation and, hence, plays R . If 1 plays aggressive then player 2 can rationalize both options and plays N . In the only subgame perfect equilibrium, player 1 is pleasant and 2 reciprocates (see Rabin (1993), Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) for alternative models of reciprocity). We also refer the reader to Spiegler (2002) and (2004) for game-theoretic models where players must justify their chosen actions.

These examples show that rationalization theory can accommodate seemingly non self-interested behavior in strategic settings. However, a general introduction of rationalization into game theory would require extending the theory to allow for choice over lotteries. This is beyond the scope of this paper.

7.3. Proofs

Proof of the preliminary result. Given a set of \mathcal{R} of rationales, assume that $x \in B \subseteq \tilde{B}$ and $x \in \tau^{\mathcal{R}}(\tilde{B})$. Then, by definition, there is some $R_i \in \mathcal{R}$ such that $x R_i y$ for all $y \in \tilde{B}$, $y \neq x$. Hence, $x R_i y$ for all $y \in B$, $y \neq x$. So, $x \in \tau^{\mathcal{R}}(B)$. It follows that 3.1 holds. Now assume that a psychological constraint function τ satisfies 3.1. Then, for each issue $B \in \mathcal{B}$ and alternative $x \in \tau(B)$, let $R_{B,x}$ be the defined by $x R_{B,x} y$ for any $y \in B$, $y \neq x$. So, $x R_{B,x} y$ if and only if $x \in \tau(B)$, $y \in B$, and $y \neq x$. Let \mathcal{R} be the set of all rationales $R_{B,x}$ such that $B \in \mathcal{B}$ and $x \in \tau(B)$. Let $\tau^{\mathcal{R}}$ be the psychological constraint function determined by \mathcal{R} . Fix any issue $B \in \mathcal{B}$. Assume that $x \in \tau(B)$. Then, by definition, x is rationalized by $R_{B,x} \in \mathcal{R}$. So, $x \in \tau^{\mathcal{R}}(B)$. Now assume that $x \in \tau^{\mathcal{R}}(B)$. So, $x \in B$ and there exists an issue \tilde{B} such that $x R_{\tilde{B},x} y$ for any $y \in B$, $y \neq x$. By definition, $x R_{\tilde{B},x} y$ if and only if $x \in \tau(\tilde{B})$, $y \in \tilde{B}$, and $y \neq x$. So, $x \in \tau(\tilde{B})$. By 3.1, $x \in \tau(B)$. ■

Given an issue B , let \mathcal{B}^B be all super-issues B^* of B such that the pair (B, B^*) violates WARP. Given a choice function C and a set $\mathcal{A} = \{(y_i, B_i); y_i \in B_i \ i = 1, \dots, n\}$, let $R^{C, \mathcal{A}}$ be the binary relation defined in section 4. That is, $x R^{C, \mathcal{A}} y$ if and only if

$$\begin{aligned} x &= C(B) \text{ and } y = C(B^*) \text{ for some pair } (B, B^*) \text{ of nested issues that violates WARP or} \\ x &= C(B) \text{ and for some } (y_i, B_i) \in \mathcal{A}, y = y_i \in B \subseteq B_i. \end{aligned}$$

Let $\tau^{C,\mathcal{A}}$ be a psychological constraint function defined by $\tau(B) =$

$$\{C(B); C(B^*) \text{ for any } B^* \in \mathcal{B}^B; y_i \text{ for any } (y_i, B_i) \in \mathcal{A}, y = y_i \in B \subseteq B_i\}.$$

By definition,

$$C(B) R^{C,\mathcal{A}} y \text{ for any } y \in \tau^{C,\mathcal{A}}(B), y \neq C(B) \quad (7.1)$$

In addition, if $B \subseteq \tilde{B}$ then

$$\tau^{C,\mathcal{A}}(\tilde{B}) \cap B \subseteq \tau^{C,\mathcal{A}}(B)$$

This follows because if $z \in B$ and $z \in \tau^{C,\mathcal{A}}(\tilde{B})$ then we can assume, without loss of generality, that $z \neq C(B)$. Otherwise $z = C(B)$ and so, $z \in \tau^{C,\mathcal{A}}(B)$. We can also assume, without loss of generality, that $z \neq C(\tilde{B})$ and that $z \neq C(\hat{B})$ for any $\hat{B} \in \mathcal{B}^{\tilde{B}}$. Otherwise (B, \tilde{B}) or (B, \hat{B}) is a pair of nested issues that violates WARP and in either case, $z \in \tau^{C,\mathcal{A}}(B)$. Hence, it follows from $z \in \tau^{C,\mathcal{A}}(\tilde{B})$ that for some $(y_i, B_i) \in \mathcal{A}$, $z = y_i \in \tilde{B} \subseteq B_i$. So, $z = y_i \in B \subseteq \tilde{B} \subseteq B_i$. Hence, $z \in \tau^{C,\mathcal{A}}(B)$. So, $\tau^{C,\mathcal{A}} \in \Upsilon^{\mathcal{A}}$.

Lemma 1. *If $(R, \tau) \in \mathcal{P}\mathbf{x}\Upsilon^{\mathcal{A}}$ underlies C then $x R^{C,\mathcal{A}} y \implies x R y$*

Proof : Assume that $x = C(B)$ and $y = C(B^*)$ for some pair (B, B^*) of nested issues that violates WARP. Then, $y \in \tau(B^*)$ (because $y = C(B^*)$) and $y \in B$ (because $C(B^*) \in B$). So, by 3.1, $y \in \tau(B)$. Hence, $x R y$ (because (R, τ) underlies C). Now assume that $x = C(B)$ and for some $(y_i, B_i) \in \mathcal{A}$, $y = y_i \in B \subseteq B_i$. So, $y_i \in \tau(B_i)$ (because $\tau \in \Upsilon^{\mathcal{A}}$). By 3.1, $y_i \in \tau(B)$. Hence, $x R y = y_i$. ■

Proposition 7. *A choice function C consistent with \mathcal{A} -rationalization theory $\mathcal{P}\mathbf{x}\Upsilon^{\mathcal{A}}$ if and only if $R^{C,\mathcal{A}}$ is asymmetric. A choice function C is consistent with \mathcal{A} -rationalization order theory $\mathcal{P}^o\mathbf{x}\Upsilon^{\mathcal{A}}$ if and only if $R^{C,\mathcal{A}}$ is acyclic.*

Proof : Assume that a choice function C is consistent with \mathcal{A} -rationalization theory $\mathcal{P}\mathbf{x}\Upsilon^{\mathcal{A}}$. Let $(R, \tau) \in \mathcal{P}\mathbf{x}\Upsilon^{\mathcal{A}}$ be a model that underlies C . Assume, by contradiction, that $R^{C,\mathcal{A}}$ is not asymmetric. Then, for some $x \neq y$, $x R^{C,\mathcal{A}} y$ and $y R^{C,\mathcal{A}} x$. By Lemma 1, $x R y$ and $y R x$. This contradicts, $R \in \mathcal{P}$. Now assume that $R^{C,\mathcal{A}}$ is asymmetric. Then, by 7.1, $(R^{C,\mathcal{A}}, \tau^{C,\mathcal{A}}) \in \mathcal{P}\mathbf{x}\Upsilon^{\mathcal{A}}$ underlies C .

Now assume that a choice function C is consistent with order \mathcal{A} -rationalization theory $\mathcal{P}^o\mathbf{x}\Upsilon^{\mathcal{A}}$. Let $(R, \tau) \in \mathcal{P}^o\mathbf{x}\Upsilon^{\mathcal{A}}$ be a model that underlies C . Assume, by contradiction, that $R^{C, \mathcal{A}}$ is cyclic. Then, Lemma 1, R is also cyclic. This contradicts, $R \in \mathcal{P}^o$. Now assume that $R^{C, \mathcal{A}}$ is acyclic. By topological ordering, $R^{C, \mathcal{A}}$ may be extended (not necessarily uniquely) to an order (see Corman et al. (2001, pp.549–552)). Let \bar{R} be an arbitrary order that extends $R^{C, \mathcal{A}}$. Then, by 7.1, $(\bar{R}, \tau^{C, \mathcal{A}}) \in \mathcal{P}^o\mathbf{x}\Upsilon^{\mathcal{A}}$ underlies C . ■

Proof of Proposition 1. Let $(R, \tau) \in \mathcal{P}\mathbf{x}\Upsilon^1$ be a model that underlies C . Note that $\Upsilon^1 = \Upsilon^{\mathcal{A}}$ where $\mathcal{A} = \emptyset$. So, if (B, B^*) is a pair of nested issues that violates WARP then $C(B) R^{C, \emptyset} C(B^*)$ and, by Lemma 1, $C(B) R C(B^*)$. So, Bob infers Dee prefers $C(B)$ to $C(B^*)$.

If C is consistent with rationalization theory $\mathcal{P}\mathbf{x}\Upsilon^1$ then, by Lemma 2, $R^{C, \emptyset}$ is asymmetric. Hence, $(R^{C, \emptyset}, \tau^{C, \emptyset}) \in \mathcal{P}\mathbf{x}\Upsilon^1$ underlies C . If there exists no pair of nested issues (B, B^*) that violates WARP such that $C(B) = x$ and $C(B^*) = y$, $x \neq y$, then it is *not* the case that $x R^{C, \emptyset} y$. Thus, consider the binary relation \bar{R} such that $y \bar{R} x$ and for all other pairs of alternatives \bar{R} is identical to $R^{C, \emptyset}$. Then, \bar{R} is still asymmetric and $(\bar{R}, \tau^{C, \emptyset}) \in \mathcal{P}\mathbf{x}\Upsilon^1$ still underlies C . ■

Proof of Proposition 2 : Let (R, τ) underlie C . Fix an issue $B \in \mathcal{B}$ and an alternative $z \in \tau(B)$. Now, $C(B) \in \tau(B)$ and $z \in \tau(B)$ implies that $\{C(B), z\} \subseteq B \cap \tau(B)$. Therefore, $\tau\{C(B), z\} = \{C(B), z\}$. Since $C(B) R z$ (because (R, τ) underlies C), it must be the case that $C(\{C(B), z\}) = C(B)$. Thus, $C(B) \tilde{R}^C z$. ■

Proof of Proposition 3 : Assume that $(R, \tau) \in \mathcal{P}\mathbf{x}\Upsilon^1$ underlies C and is not dominated. Then, for every pair $\{x, y\} \subset A$, $\tau\{x, y\} = \{x, y\}$. To see this assume, by contradiction, that for some pair of alternatives $\{x, y\}$, $\tau\{x, y\} = \{x\}$. Let τ' be such that $\tau'\{x, y\} = \{x, y\}$ and $\tau' = \tau$ for all other issues. Clearly, $\tau' \in \Upsilon^1$ because $\tau \in \Upsilon^1$ and $\{x, y\}$ has no non-trivial sub-issues (issues with more than one alternative). By assumption \tilde{R}^C (defined in the main text) is complete and acyclical and so is an order. We now show that $(\tilde{R}^C, \tau') \in \mathcal{P}^o\mathbf{x}\Upsilon^1$ underlies C .

Let $B \neq \{x, y\}$ be an issue. Let $z \in \tau'(B) = \tau(B)$, $z \neq C(B)$. Note that $\{C(B), z\} \subseteq B$ and $C(B) \in \tau(B) = \tau'(B)$. So, $\{C(B), z\} \subseteq B \cap \tau'(B)$ and $\{C(B), z\} \subseteq B \cap \tau(B)$. Hence, $\tau(\{C(B), z\}) = \tau'(\{C(B), z\}) = \{C(B), z\}$. It follows that $C(B) R z$ (because $z \in \tau(\{C(B), z\})$ and (R, τ) underlies C). Hence, $C(\{C(B), z\}) = C(B)$ (because (R, τ) underlies C). By definition, $C(B) \tilde{R}^C z$. Moreover, $C(\{x, y\}) = x$ (because $\tau\{x, y\} = \{x\}$). So, $x \tilde{R}^C y$. Hence, $(\tilde{R}^C, \tau') \in \mathcal{P}^o\mathbf{x}\Upsilon^1$ underlies C and (R, τ) is

dominated by (\tilde{R}^C, τ') . Thus, for every pair of alternatives $\{x, y\}$, $\tau\{x, y\} = \{x, y\}$. Given that (R, τ) underlies C it now follows that $R = \tilde{R}^C$. ■

Proof of Proposition 4. The proof of proposition 4 mimics the proof of proposition 1. Let $(R, \tau) \in \mathcal{P}_x\Upsilon^A$ be a model that underlies C . So, if either condition 1 or 2 holds then $x R^{C,A} y$ and, by Lemma 1, $x R y$. Now assume that C is consistent with \mathcal{A} -rationalization theory $\mathcal{P}_x\Upsilon^A$. Then, by Lemma 2, $R^{C,A}$ is asymmetric. Hence, $(R^{C,A}, \tau^{C,\emptyset}) \in \mathcal{P}_x\Upsilon^A$ underlies C . If neither condition 1 nor condition 2 holds then it is *not* the case that $x R^{C,A} y$. Thus, consider the binary relation \bar{R} such that $y \bar{R} x$ and for all other pairs of alternatives \bar{R} is identical to $R^{C,A}$. Then, \bar{R} is still asymmetric and $(\bar{R}, \tau^{C,A}) \in \mathcal{P}_x\Upsilon^A$ still underlies C . ■

Proof of Proposition 5. Let $(R, \tau) \in \mathcal{P}^o_x\Upsilon^A$ be a model that underlies C . If there exists a chain z_{i+1} , $i = 0, \dots, k-1$, such that z_i is revealed preferred to z_{i+1} by \mathcal{A} -rationalization theory then $z_i R z_{i+1}$, $i = 0, \dots, k-1$, which implies (because R is an order) that $z_1 R z_k$. Now assume that C is consistent with \mathcal{A} -rationalization order theory $\mathcal{P}^o_x\Upsilon^A$. By Lemma 2, $R^{C,A}$ is acyclic. Now assume that there is no chain z_{i+1} , $i = 0, \dots, k-1$, such that z_i is revealed preferred to z_{i+1} by \mathcal{A} -rationalization theory. Hence, it is *not* the case that $z_1 R^{C,A} z_k$. Thus, consider the binary relation \bar{R} such that $z_k \bar{R} z_1$ and for all other pairs of alternatives \bar{R} is identical to $R^{C,A}$. Then, \bar{R} is still acyclic and, hence, can be extended to an order $\hat{R} \in \mathcal{P}^o$. Given that \hat{R} extends $R^{C,A}$, and $(R^{C,A}, \tau^{C,A}) \in \mathcal{P}_x\Upsilon^A$ still underlies C , $(\hat{R}, \tau^{C,A}) \in \mathcal{P}^o_x\Upsilon^A$ still underlies C . ■

Proof of Proposition 6: Assume that WWARP does not hold. Then let $x \neq y$, $\{x, y\} \subseteq B \subseteq \bar{B}$ be such that $C(\bar{B}) = C(\{x, y\}) = x$ and $C(B) = y$. Then, $(\{x, y\}, B)$ is a pair of nested issues that violates WARP and (B, \bar{B}) is also a pair of nested issues that violates WARP. But $C(\bar{B}) = C(\{x, y\}) = x$. Hence, $(\{x, y\}, B)$ and (B, \bar{B}) are reversed. Thus, the irreversibility axiom does not hold.

Now assume that the irreversibility axiom does not hold. Consider the two pairs (B_1, B_1^*) and (B_2, B_2^*) of reversed nested issues that violate WARP. Let $y = C(B_1) = C(B_2^*)$ and $x = C(B_1^*) = C(B_2)$. Then, $x \neq y$, $\{x, y\} \subseteq B_1 \subseteq B_1^*$ and $\{x, y\} \subseteq B_2 \subseteq B_2^*$ ($x \in B_1$ because $x = C(B_1^*) \in B_1$ and $y \in B_1$ because $y = C(B_1) \in B_1$). So, $\{x, y\} \subseteq B_1$. The argument for $\{x, y\} \subseteq B_2$ is analogous). Now assume that $C(\{x, y\}) = x$. Then, $\{x, y\} \subseteq B_1 \subseteq B_1^*$, $C(B_1^*) = x$ and $C(B_1) = y$. So, WWARP does not hold. On the other hand if $C(\{x, y\}) = y$ then $\{x, y\} \subseteq B_2 \subseteq B_2^*$, $C(B_2^*) = y$ and $C(B_2) = x$. Thus, WWARP does not hold. ■

References

- [1] Akerlof, G. and W. Dickens (1982) “The Economic Consequences of Cognitive Dissonance,” *American Economic Review*, 72 (3), 307-319.
- [2] Achen, A. and L. Bartels (2006) “It Feels Like We’re Thinking: The Rationalizing Voter and Electoral Democracy,” mimeo.
- [3] Ambrus, A., and K. Rozen. (2008) "Revealed conflicting preferences". Working paper Harvard University.
- [4] Andreoni, J. (1989) “Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence,” *Journal of Political Economy*, 97, 1447-1458.
- [5] Andreoni, J. (1990) “Impure Altruism and Donations to Public Goods. A Theory of Warm-Glow Giving,” *Economic Journal*, 100, 464-477.
- [6] Andreoni, J. (2006) “Philanthropy” mimeo Wisconsin University.
- [7] Dan Ariely, D. and N. Mazar (2006) “Dishonesty in Everyday Life and its Policy Implications” *Journal of Public Policy and Marketing*, 25-1, 117-126.
- [8] Becker, G. (1974) “A Theory of Social Interactions,” *Journal of Political Economy*, 82, 1063–1093.
- [9] Bénabou, R. and J. Tirole (2002) “Self-Confidence and Personal Motivation,” *Quarterly Journal of Economics*, 117 (3), 871–915.
- [10] Berger, P. and G. Smith (1997) “The Effect of Direct Mail Framing Strategies and Segmentation Variables on University Fundraising Performance,” *Journal of Direct Marketing*, 2 (1), 30-43.
- [11] Bernheim, D. (1984) “Rationalizable Strategic Behavior,” *Econometrica*, 52 (4), 1007–1028 .
- [12] Bernheim, D. and A. Rangel (2007) “Toward Choice-Theoretic Foundations for Behavioral Welfare Economics,” *American Economic Review, papers and proceedings*, 97, 464–470.

- [13] Blow, L., M. Browning, and I. Crawford (2008) "Revealed Preference Analysis of Characteristics Models," *Review of Economic Studies*, 75, 371–389.
- [14] Bolton, G. and A. Ockenfels (2000) "ERC A Theory of Equity, Reciprocity and Competition," *American Economic Review*, 90 (1), 166–193.
- [15] Carillo, J. and T. Mariotti (2000) "Strategic Ignorance as a Self-Discipline Device," *Review of Economic Studies*, 67 (3), 529–544.
- [16] Chambers. C. and T. Hayashi (2008) "Choice and Individual Welfare," mimeo Caltech.
- [17] Chen, K. (2008) "Rationalization and Cognitive Dissonance: Do Choices Affect or Reflect Preferences," mimeo.
- [18] Cormen, T., C. Leiserson, R. Rivest, and C., Stein. (2001) "Introduction to Algorithms." MIT Press and McGraw-Hill. Second Edition.
- [19] Dahl, G. and M. Ransom (1999) "Does Where You Stand Depend on Where You Sit? Tithing Donations and Self-Serving Beliefs," *American Economic Review*, 89 (4), 703–727.
- [20] Dekel, E., B. Lipman, and A. Rustichini (2001) "Representing Preferences with a Unique Subjective State Space," *Econometrica* 69 (4) , 891–934.
- [21] Doyle, J., D. O'Connor, G. Reynolds, and P. Bottomley (1999) "The Robustness of the Asymmetrically Dominated Effect: Buying Frames, Phantom Alternatives, and In-Store Purchases," *Psychology and Marketing*, 16, 225-243.
- [22] Easley, D. and A. Rustichini (1999) "Choice Without Beliefs," *Econometrica* 67 (5), 1157–1184.
- [23] Eliaz, K. and E. Ok (2006) "Indifference or indecisiveness? Choice-theoretic Foundations of Incomplete Preferences" *Games and Economic Behavior*, 56, 61–86.
- [24] Feddersen, T. (2004) "Rational Choice Theory and the Paradox of Voting." *Journal of Economic Perspectives*, 18 (1), 99-112.

- [25] Fehr, E., and K. Schmidt (1999) "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114, 817–68.
- [26] Fishburn, P. and I. LaValle (1988) "Context-Dependent Choice with Nonlinear and Nontransitive Preferences," *Econometrica*, 56 (5), 1221-1239.
- [27] Fudenberg, D. and D. Levine (2006) "A Dual-Self Model of Impulse Control," *American Economic Review*, 96 (5), 1449-1476.
- [28] Gale, D. and Mas-Colell, A. (1975) "An equilibrium Existence Theorem for a General Model without Ordered preferences" *Journal of Mathematical Economics* 2, 9–15.
- [29] Goldstein, N.J., and Cialdini, R.B. (2007) "Using social norms as a lever of social influence." In A. Praktanis (Ed.), *The science of social influence: Advances and future progress*. Philadelphia: Psychology Press.
- [30] Green, J., and D. Hojman. (2007) "Choice, Rationality and Welfare Measurement". Harvard University. Working Paper.
- [31] Gul, F. and W. Pesendorfer (2001) "Temptation and Self-Control," *Econometrica* 69 (6) , 1403–1435.
- [32] Gul, F. and W. Pesendorfer (2004) "Self-Control and the Theory of Consumption," *Econometrica* 72 (1) , 119–158.
- [33] Gul, F. and W. Pesendorfer (2005) "The Revealed Preference Theory of Changing Tastes," *Review of Economic Studies* 72 (2) , 429–448.
- [34] Gul, F. and W. Pesendorfer (2005) "The Case for Mindless Economics," mimeo.
- [35] Hubler, J., J. Payne, and C. Puto (1982) "Adding Asymmetrically Dominated Alternatives; Violation of Regularity and the Similarity Hypothesis," *Journal of Consumer Research*, 9, 90-98.
- [36] Iyengar, S., and M. Lepper (2000) "When Choice is Demotivating: Can One Desire Too Much of a Good Thing?" *Journal of Personality and Social Psychology*, 79, 995-1006.

- [37] Jones, E. (1908) “Rationalisation in Every-day Life,” *Journal of Abnormal Psychology*, 161-169.
- [38] Kalai, G., A. Rubinstein, and R. Spiegler (2002) “Rationalizing Choice Functions by Multiple Rationales,” *Econometrica*, 70 (6), 2481-2488.
- [39] Maccheroni, F., M. Marinacci and A. Rustichini (2006) “Ambiguity Aversion, Robustness, and the Variational Representation of Preferences,” *Econometrica*, 74 (6), 1447-1498.
- [40] Manzini, P. and M. Mariotti (2007) “Sequentially Rationalizable Choice” *American Economic Review* **97**-5, 1824-1839.
- [41] Manzini, P. and M. Mariotti. (2009). “Boundedly Rational Choice, Cycles, and Menu Effects: Theory and Experimental Evidence.” mimeo, Queen Mary, University of London.
- [42] Masatlioglu, Y. and D. Nakajima (2007) “A Theory of Choice by Elimination,” mimeo.
- [43] Masatlioglu, Y. and E. Ok (2007) “Status Quo Bias and Reference-Dependent Procedural Decision Making,” mimeo.
- [44] Mas-Colell, A. (1974) “An equilibrium Existence Theorem without Complete or Transitive Preferences,” *Journal of Mathematical Economics*, 1, 237–246.
- [45] May, K (1954) “Intransitivity, Utility, and the Aggregation of Preference Patterns,” *Econometrica*, 22 (1), 1-13.
- [46] Mill J.S. (1860): *On Liberty*, P.F. Collier & Son, London.
- [47] Moulin, H. (1985) “Choice Functions over a Finite Set: A Summary,” *Social Choice and Welfare*, 2, 147-160.
- [48] Mullainathan, S., and R. Thaler (2000) “Behavioral Economics,” mimeo.
- [49] Ok, E., P. Ortoleva, and G. Riella (2008) “Rational Choice with Endogenous Reference Points,” mimeo.
- [50] Pearce D. (1984) “Rationalizable Strategic Behavior and the Problem of Perfection,” *Econometrica*, 52, (4), 1029-1050.

- [51] Rabin, M. (1993) “Incorporating Fairness into Game Theory and Economics.” *American Economic Review*, 83 (4), 1281–1302.
- [52] Rabin, M. (1995) “Moral Preferences, Moral Constraints, and Self-Serving Biases,” mimeo.
- [53] Rabin, M. (1998) “Psychology and Economics,” *Journal of Economic Literature*, Vol. 36 (1), 11-46.
- [54] Roth, A. (2007) “Repugnance as Constraints on Markets,” mimeo Harvard University.
- [55] Salant, Y. and A. Rubinstein (2006) “Two Comments on the Principle of Revealed Preferences,” mimeo.
- [56] Salant, Y. and A. Rubinstein (2006a) “A Model of Choice from Lists,” *Theoretical Economics*, 1, 3-17.
- [57] Samuelson, P. (1938a) “A Note on the Pure Theory of Consumer’s Behavior,” *Economica*, 5 (17), 61-71.
- [58] Samuelson, P. (1938b) “The Empirical Implications of Utility Analysis,” *Econometrica*, 6 (4), 344-356.
- [59] Sen, A. (1997) “Maximization and the Act of Choice,” *Econometrica*, 65 (4), 745-779.
- [60] Simon, H. (1959) “Theories of Decision-Making in Economics and Behavioral Science,” *American Economic Review*, 49 (3), 253-283.
- [61] Simonson, I. (1989) “Choice Based on Reasons: The Case of Attraction and Compromise Effects,” *The Journal of Consumer Research*, 16 (2), 158-174.
- [62] Simonson, I. and A. Tversky (1992) “Choice in Context: Tradeoff Contrast and Extremeness Aversion,” *Journal of Marketing Research*, 29, 281-295.
- [63] Simonson, I and A. Tversky (1993) “Context-Dependent Preferences,” *Management Science*, 39 (10), 1179-1189.
- [64] Shafer, W. and H. Sonnenschein (1975) “Equilibrium in Abstract Economies without Ordered Preferences,” *Journal of Mathematical Economics*, 2, 345-348.

- [65] Shafir, E., I. Simonson, and A. Tversky (1993) “Reason-based Choice,” *Cognition* 49, 11-36.
- [66] Sprumont, Y. (2000) “On the Testable Implications of Collective Choice Theories,” *Journal of Economic Theory*, 93, 205-232.
- [67] Spiegel, R. (2002) “Equilibrium in Justifiable Strategies: A Model of Reason-Based Choice in Extensive-Form Games” (2002), *Review of Economic Studies* 69, 691-706.
- [68] Spiegel, R. (2004) “Simplicity of Beliefs and Delay Tactics in a Concession Game,” *Games and Economic Behavior* 47 (1), 200-220.
- [69] Snyder, M., R. Kleck, A. Strenta, and S. Mentzer (1979) “Avoidance of the Handicapped: An Attributional Ambiguity Analysis,” *Journal of Personality and Social Psychology*, 37 (12), 2297-2306.
- [70] Thaler R. and C. Sunstein (2003): “Libertarian Paternalism”, *American Economic Review* 93: 175-179
- [71] Tversky, A. and E. Shafir (1992) “Choice Under Conflict: The Dynamics of Deferred Decision,” *Psychological Science*, 3 (6), 358-361.
- [72] von Hippel, W., J. Lakin, and R. Shakarchi (2005) “Individual Differences in Motivated Social Cognition: The Case of Self-Serving Information Processing,” *Personality and Social Psychology Bulletin*, 31 (10), 1347-1357.