Rationalization*

Vadim Cherepanov, Timothy Feddersen and Alvaro Sandroni

November 17, 2010

Abstract

In 1908 the Welsh neurologist and psychoanlayst 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 this model a decision maker selects her preferred alternative from among those that she can rationalize. We show that this theory is testable and broadens standard economics in a natural way. In particular, choice may fully reveal preferences even when standard theory does not apply. Rationalization theory allows for a new way to interpret speech and can reveal hidden preferences for discrimination. In addition, rationalization theory can aid data analysis by characterizing the conditions necessary to reach a given conclusion. Finally, rationalization theory can be used to understand behavioral changes in the absence of changes in preferences, incentives and opportunities.

*We thank Eddie Dekel, Jennifer Jordan, Paola Manzini, Marco Mariotti, Herve Moulin, Efe Ok, Nicola Persico, Charlie Plott, Scott Presti, Yuval Salant, Yves Sprumont, Marciano Siniscalchi for useful comments. Vadim Cherepanov [vadimch@sas.upenn.edu] may be reached at the Department of Economics, University of Pennsylvania. Alvaro Sandroni [sandroni@kellogg.northwestern.edu] and Tim Feddersen [tfed@kellogg.northwestern.edu] may be reached at the Kellogg School of Management, Northwestern University. Sandroni gratefully acknowledges financial support from the National Science Foundation. All error are ours.

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 Sigmund Freud with the critical insight "that a number of mental processes owe their origin to causes unknown to and unsuspected by the individual" he writes that *rationalization* occurs because people feel "*a necessity to provide an explanation*" (Jones (1908)).

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 and accept implausible explanations for their behavior. However, the phenomena of rationalization can only influence choice if the inability to rationalize constrains behavior.

While standard economics accommodates physical constraints, *psychological constraints* have not yet received much attention in the academic economics literature. An exception is Roth (2007) who lists potentially beneficial practices that have been deemed repugnant and banned. Some examples include banning the human consumption of horse meat (illegal in California), selling pollution permits and markets for human organs. These examples illustrate that psychological constraints on choice are sometimes binding.

Rationalization is economically relevant in seemingly unrelated settings. For example, Aronson and Pratkanis (2001) describe several marketing techniques that exploit consumers' need to rationalize choice. The Statement of Accounting Standards (SAS99) identifies the ability to rationalize as a central risk factor in fraud: "[t]hose involved in a fraud are able to rationalize a fraudulent act as being consistent with their personal code of ethics." To appreciate the potential significance of rationalization one need only consider the social cost of fraud. Buckoff (2001) reports that employee fraud alone costs employers $400 billion or about 6% of their revenues.

We seek to better understand the underlying logic of rationalization by developing a formal model. Our main premise is that agents choose according to their preferences,

[1] David 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."

as in standard economics, but face potentially unobservable psychological constraints. For example, a manager may have the opportunity and incentive to commit fraud but, absent a convincing rationale that legitimizes fraud, will choose not to do so.

We model a decision maker (Dee) who has preferences over alternatives but, unlike the standard theory, also has a set of *rationales* (modeled as 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 subjectively appealing rationale that can justify that choice. Rationalization, like standard economic theory, is a constrained optimization process. However, unlike standard theory, the constraints on choice need not be observable.

Consider the following scenario. Given the choice between work (w) and a movie (m) Dee chooses the movie. However, when Dee is given a third alternative of visiting a relative in the hospital (h) she stays at work. Dee's choice of m from the set $\{w, m\}$ and w from the set $\{w, m, h\}$ violates *the Weak Axiom of Revealed Preferences (WARP)* (see Samuelson (1938)).[2] In standard economics violations of WARP are *anomalous* because they lead to contradictory inferences of preferences, but rationalization theory can accommodate this behavior.

Suppose Dee 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 w is ranked above m and h. Under rationale 2 work is not pressing and h is ranked above m and m above w. Dee chooses the movie over work because she prefers it 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) but can rationalize her second choice of work (by rationale 1).

A wide variety of behavioral anomalies can be accommodated by rationalization theory. Indeed, Dee's preferences can be captured by a stable, single order even if observed choices are cyclic. Nevertheless, rationalization theory is testable. A simple and known axiom akin to WARP fully characterizes the empirical content of rationalization theory.

Rationalization theory delivers a new way to interpret behavior including speech. Consider the effort and passion spent on seemingly abstract debates over social values

[2]WARP states that if x is chosen over y then y is not chosen from any set of alternatives that includes x. We use the acronym WARP to replace "the Weak Axiom of Revealed Preference."

(e.g., with respect to marriage, relations between adults and children, and attitudes towards stigmatized behavior). In standard theory, speech has no effect on choice unless it changes preferences (perhaps by altering beliefs). In rationalization theory, psychological constraints may be created or removed when rationales are legitimized or delegitimized. As a result behavior may change even if preferences, opportunities and incentives remain unchanged. In a world of rationalizers, speech becomes contentious because it changes behavior. Rationalization theory also provides a natural framework to interpret contradictory speech; the consequences of legal rulings; and demotivating aspects of choice. We will expand on these applications below.

A central contribution of this paper is to demonstrate that rationalization theory can shed new light on empirical work in a variety of disciplines. For example, a large literature in social psychology investigates the way in which people respond to stigmatized groups. In a well-known experiment, Snyder et al. (1979) allow subjects to choose whether to watch a movie alone or with someone in a wheelchair. In one treatment subjects disproportionally choose to watch a movie with a person in a wheelchair rather than watching the same movie alone. In a second treatment, when the movies are different, subjects disproportionately choose to watch a movie alone rather than with the handicapped person. The experiment was designed to rule out actual preferences between the movies as an explanation for behavior. The interpretation of Snyder et al. is that many subjects want to avoid the handicapped. In the first treatment subjects are psychologically constrained to watch the movie with the handicapped person because to do otherwise would require subjects to reveal (perhaps only to themselves) handicapped aversion.

Rationalization theory captures the behavior in Synder's study in a way that is consistent with their interpretation of the data: Dee prefers to see the movie alone, but cannot rationalize doing so when the movies are the same. In the first treatment Dee acts against her preferences because of a psychological constraint while in the second treatment the constraint is relaxed. She can rationalize watching the second movie alone by telling herself that she prefers that movie. The implication of the Synder study is that legitimizing socially undesirable behavior (such as discrimination) may remove psychological constraints resulting in actual undesirable conduct.

In order to show handicapped aversion it is necessary to show that individuals in the first treatment prefer to avoid the handicapped even though their behavior seems to contradict that conclusion. Our model reveals that such a demonstration

is more difficult than it might seem. The problem is that the observed behavior is also consistent with preferences and psychological constraints that do not require handicapped aversion.

The value of a formal model is that it reveals precisely what is necessary and sufficient to reach a given conclusion. In the paper we show that the conclusion of handicapped aversion requires two conditions that seem contradictory in both economics and psychology: first, cyclic choice behavior must be observed and, second, decision makers must have ordered preferences over outcomes. In addition, the handicapped aversion result also requires the assumption that subjects can rationalize seeing a movie with the handicapped person instead of watching a different movie alone. This assumption is an example of a *permissibility assumption.*

Permissibility assumptions stipulate that Dee can rationalize a given option in a given set of alternatives. The assumption that Dee cannot rationalize a choice is called a *impermissibility assumption.* We call both permissibility and impermissibility assumptions *contextual* to emphasize their dependence on the specific context of choice. We fully characterize inferences that can be made (from choice) with and without contextual assumptions. In addition, we also show that only permissibility assumptions help identify preferences while impermissibility assumptions do not. These results reveal necessary and sufficient conditions for inferring preferences and constraints from choice.

A preference order is *identifiable* from observed choices if there exist any set of contextual assumptions that pin it down uniquely. We show that a preference order is identifiable if and only if it satisfies the *minimum constraint principle* i.e., accommodates Dee's observed choices while imposing minimal constraints on what she might choose. In particular, the only preferences that can be identified by data satisfy the minimum constraint principle.

We define the *minimum constraint theory of rationalization* as the set of rationalization models that satisfy the minimum constraint principle. If behavior is not anomalous, the minimum constraint theory of rationalization reveals the same preference order as standard economic theory. Hence, all the insights of standard economics are left undisturbed. However, the minimum constraint rationalization theory also fully reveals preferences and constraints in a wide variety of settings in which choice is anomalous. If Dee's binary choices are acyclical her revealed preferences are given by her binary choices and all constraints are revealed. Dee's preferences and con-

straints are sometimes fully revealed even when observed choice is cyclic. Hence, the minimum constraint theory of rationalization broadens standard economics in a natural way: it coincides with standard economics when standard theory holds and also allows for precise inferences in several case where standard economics makes contradictory inferences. These results show that psychology is more aligned with standard economics than commonly perceived. Moreover, these results can be seen as a first step towards welfare analysis based on psychological ideas.

The organization of the paper is as follows. Section 2 provides a brief literature review. Section 3 formalizes the idea of rationalization. Section 4 shows results on revealed preferences and some of the implications of these results for empirical work. Section 5 introduces the minimum constraint principle. Section 6 characterizes the empirical content of rationalization theory. Section 7 shows the implications of rationalization theory for behavior modification, political debate, interpretation of contradictory speech, incorporation of rationalization theory in economics, and legal studies. Section 8 provides a conclusion. Proofs are in the appendix.

2. Related Literature

A growing literature focuses on conflicting motivations. See, among many contributions, Ambrus and Rozen (2008), Bernheim and Rangel (2009), Chambers and Hayashi (2008), Clippel and Eliaz (2009), Dietrich and List (2010), Fudenberg and Levine (2006), Green and Hojman (2007), Gul and Pesendorfer (2005), Kalai, Rubinstein and Spiegler (2002), Heller (2009), Lehrer and Teper (2009), Manzini and Mariotti (2007, 2009), Ok, Ortoleva, and Riella (2008), Salant and Rubinstein (2006, 2006a). While these models do not formalize the idea of rationalization as we define it, they can accommodate behavioral anomalies. In section 7 we discuss how to differentiate among theories that are consistent with the same observed behavior.

The word "rationalizability" is used in game theory (see Bernheim (1984), Pearce (1984), Sprumont (2000)) quite differently from us, but share a common idea that actions can be taken when justified. The word "rationalization" is also used differently in cognitive dissonance theory. The basic claim is that people devalue rejected choices and valorize 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), Carrillo and Mariotti

(2000) and Bénabou and Tirole (2002)). A large literature also 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.

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 $C : \mathcal{B} \longrightarrow A$ 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 decision maker, named Dee, makes the choices given by C.

A *preference* P is an asymmetric binary relation on A. A transitive, complete preference is an *order*. As usual, $x \ P \ y$ denotes that x is $P-$preferred to y. A *psychological constraint function* is a mapping $\psi : \mathcal{B} \longrightarrow \mathcal{B}$ such that $\varnothing \neq \psi(B) \subseteq B$ for every issue $B \in \mathcal{B}$. An option $x \in \psi(B)$ is *psychologically feasible in issue B*. Dee chooses the option she prefers among those that are psychologically feasible. We refer to psychologically feasible options as *permissible* and psychologically infeasible options as *impermissible*.

A *model of behavior* is a pair (P, ψ) of a preference and a psychological constraint function. A model of behavior (P, ψ) *underlies* a choice function C if for any issue $B \in \mathcal{B}$, $C(B) \in \psi(B)$ and

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

So, Dee chooses as if she solves an optimization problem with (possibly) psychological constraints. The standard theory of choice is a special case without binding psychological constraints, i.e., $\psi(B) = B$. What differentiates rationalization theory and standard economics is that psychological constraints may be unobservable.

If unobservable constraints are not structured then any pattern of choice can occur. That is, any choice of x in issue B can be accommodated by a model in which Dee's choice is dictated entirely by her constraints, i.e., $\psi(B) = \{x\}$. So, a testable theory must put restrictions on psychological constraints. Rationalization imposes a logical structure on psychological constraints that allows for testability.

Let Ψ be the set of all psychological constraint functions. Let $\mathcal{P}$ be the set of

all preferences. A *theory of behavior* is a subset $\hat{\mathcal{P}}\mathrm{x}\hat{\Psi} \subseteq \mathcal{P}\mathrm{x}\Psi$ of preferences and psychological constraint functions. So, a theory of behavior is a collection of models of choice, defined by preferences and/or psychological constraints.

We 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 explain (at least to herself) why x might be the best choice. The determination that one choice is better than another implies a comparison that can be formalized by a binary relation (not necessarily complete, transitive or asymmetric) R on A called a *rationale*. So, $x\ R\ y$ indicates that x can be rationalized over y by rationale R. This means that Dee has at least one way to tell herself that x is at least as good as y. Given an issue B, we say an alternative $x \in B$ is *rationalized* by R if and only if $x\ R\ y$ for all $y \in B$, $y \neq x$.

Let $\mathcal{R} = \{R_i,\ i = 1, ..., n\}$ be the set of all Dee's rationales. Given $\mathcal{R}$, an option $x \in B$ is *rationalizable* in B if x can be rationalized by some rationale $R_i \in \mathcal{R}$ that Dee accepts. A set of rationales $\mathcal{R}$ defines a psychological constraint function $\psi^{\mathcal{R}}$ where $\psi^{\mathcal{R}}(B)$ is the set of rationalizable options in B.

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

$$\text{if } B \subseteq B^* \text{ then } \psi(B^*) \bigcap B \subseteq \psi(B). \tag{3.1}$$

Moreover, if a psychological constraint function ψ 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 $\psi = \psi^{\mathcal{R}}$.

So, 3.1 is the key structure on psychological constraints required by rationalization. In addition, there is no loss of generality in assuming that rationales are transitive and asymmetric.[3] Informally, if Dee can rationalize an option x from a larger issue B^* then she can also rationalize it in a subset B of these alternatives. So, it becomes harder to rationalize an option when the set of alternatives grows.

Let $\Psi^{\mathcal{R}} \subseteq \Psi$ be the set of psychological constraint functions that satisfy 3.1. These are the psychological constraints implied by rationales. Let $\mathcal{P}\mathrm{x}\Psi^{\mathcal{R}}$ be the *basic theory of rationalization*. Let $\mathcal{P}^o \subseteq \mathcal{P}$ be the set of preferences orders. We define $\mathcal{P}^o\mathrm{x}\Psi^{\mathcal{R}}$ as a *theory of order rationalization*, i.e., rationalization theory with the restriction that preferences are orders.

[3] If rationales must be orders then psychological constraint functions is strictly contained in $\Psi^{\mathcal{R}}$.

In applied contexts it may be appropriate to make additional assumptions about Dee's psychological constraints. For example, a theorist (Bob) may know that Dee belongs to a religious organization that valorizes helping the needy. Alternatively, Bob might observe that Dee reveals a rationale through speech (e.g., by saying that people should help the needy). In either case Bob might feel justified in assuming that helping the needy is psychologically permissible for Dee. Alternatively, Bob may know that Dee belongs to a group that bans a certain activity e.g., eating certain types of food is taboo. Then, Bob may assume that a given choice is impermissible.

We do not claim that assumptions about psychological constraints should or will be made. Instead, we are interested in the extent to which such assumptions may help reveal preferences or reject rationalization theory.

Formally, *permissibility assumptions* are a set $\mathcal{A} = \{(y_i, B_i);\ y_i \in B_i\ i = 1, ..., n\}$ of n issues $B_i \in \mathcal{B}$ and alternatives $y_i \in B_i$ implying that $y_i \in B_i$ is psychologically feasible at B_i. Let $\Psi^{\mathcal{A}} \subseteq \Psi^{\mathcal{R}}$ be the set of all psychological constraint functions ψ that satisfy 3.1 and such that $y_i \in \psi(B_i)$, $i = 1, ..., n$. Let $\mathcal{P}\mathrm{x}\Psi^{\mathcal{A}}$ be the *theory of* $\mathcal{A}-$*rationalization* and $\mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$ be the *theory of order* $\mathcal{A}-$*rationalization.* Note that the fact that $(y, B) \notin \mathcal{A}$ means that y is not assumed to be permissible in B. It does not mean that y is assumed to be impermissible in B. We discuss the formalization of impermissibility assumptions below.

4. Revealed Preferences

In the next two sections we show how to identify preferences and constraints from choice. A choice function C is *consistent* with a theory of behavior $\hat{\mathcal{P}}\mathrm{x}\hat{\Psi}$ if some model of behavior $(P, \psi) \in \hat{\mathcal{P}}\mathrm{x}\hat{\Psi}$ underlies C. So, a choice function is consistent with a theory if the choices are produced by a model within the theory.

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

So, Bob infers that Dee prefers x over y if x ranks higher than y in Dee's preferences for every model of behavior (within Bob's theory) that underlies her choices.

Let us now consider the inferences that follow from basic rationalization theory. 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*. 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 anomalous.

Proposition 1. *Let C be a choice function consistent with Bob's basic rationalization theory $\mathcal{P}x\Psi^{\mathcal{R}}$. 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 $y = C(B^*)$.*

Proposition 1 delivers a full characterization of revealed preferences from choice under basic rationalization theory. In the appendix, we show an analogous characterization for the theory of $\mathcal{A}-$rationalization and the theory of order $\mathcal{A}-$rationalization.

Proposition 1 shows that preferences are revealed if and only if an anomaly has been observed. The intuition is as follows: a choice of y in the super-issue reveals that y is permissible in the super-issue and, therefore, in every sub-issue as well. Thus, the choice of x in the sub-issue reveals a preference for x over y (while the choice of y in the super-issue reveals that x is impermissible in the super-issue).

We now illustrate how rationalization theory can aid empirical work by clarifying the conditions under which inferences about preferences can emerge.

4.1. Handicap Aversion

As mentioned in the introduction, Snyder et. al., (1979) design an experiment intended to demonstrate handicap aversion. In the experiment, there are three alternatives: watch movie 1 alone (x); watch movie 2 alone (y); and watch movie 1 with a person in a wheelchair (z). Several subjects choose to watch the movie 1 with the handicapped rather than movie 1 alone (i.e., $\bar{C}(x, z) = z$). In addition, many subjects also choose to watch movie 2 alone rather than movie 1 with the handicapped (i.e., $\bar{C}(y, z) = y$).

Snyder et. al., claim that some subjects prefer to avoid the handicapped (i.e., prefer x to z) but, nevertheless, choose z rather than x. Their idea is that subjects can't rationalize what they prefer (to see the movie alone) when the movies are the same but they can rationalize it when the movies are different.

Rationalization theory can establish necessary and sufficient conditions for the handicap aversion hypothesis. Consider the choice between movie 1 and movie 2 (i.e., between x and y). Some choose x and some choose y. The behavior of those who choose y can be explained by standard theory without handicap aversion or psychological constraints. These subjects simply prefer y to z to x. So, consider those who choose x (i.e., $\bar{C}(x,y) = x$). Now the observed choice behavior is cyclic: x chosen over y, y chosen over z and z chosen over x. So, this cycle is necessary to show handicap aversion. However, simply observing cyclic choice is not sufficient. By proposition 1, no matter which option is chosen when all three options x, y and z are available, it does *not* follow that x is preferred to z because x was *not* chosen over z in the binary choice.

Suppose that Dee chooses to see movie 2 alone when all three options are available (i.e., $\bar{C}(x,y,z) = y$). The handicap avoidance interpretation can be captured by a rationalization model, let's call it the $S-$model, where Dee's preference order is x to y to z and the psychological constraints are $\psi\{x,z\} = \{z\}$, $\psi\{x,y,z\} = \{y,z\}$, $\psi\{B\} = B$ in all other issues. The $S-$model accommodates the choices in $\bar{C}$ and perfectly captures the intuition in Snyder et. al., because Dee prefers to avoid the handicapped (x preferred to z), but cannot rationalize her preferred choice ($\psi\{x,z\} = \{z\}$) when the movies are the same. However, consider another model, the $S'-$model, where the preference order is z to x to y; $\psi'\{y,z\} = \psi'\{x,y,z\} = \{y\}$, $\psi'\{B\} = B$ elsewhere. The $S'-$model also accommodates the choices in $\bar{C}$ even though it does not involve handicap aversion. Hence, the handicapped avoidance result cannot yet be obtained.

4.2. Preferences and Binary Choice

In this section, we show that the Synder et. al., conclusion of handicapped avoidance cannot be established by basic rationalization theory and *any* permissibility assumptions.

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

Proposition 2. *Consider a model $(P,\psi) \in \mathcal{P}x\Psi^{\mathcal{R}}$ that underlies a choice function C. Then, the model $\left(P^C,\psi\right)$ also underlies the choice function C.*

Proposition 2 shows that if a choice function is consistent with basic rationalization theory then it is always possible to accommodate Dee's choices by preferences defined

by her binary choices. So, in the handicapped avoidance example, there is a third way to accommodate the choice function $\bar{C}$: with cyclic preferences $P^{\bar{C}}$ and the psychological constraints in the $S-$model or the $S'-$model. The intuition is as follows: Dee prefers her choice $C(B)$ over any rationalizable option $z \in \psi(B)$. In a binary choice between z and $C(B)$, both options are rationalizable (they are rationalizable in the super-issue B). Thus, Dee chooses $C(B)$ over z in a binary choice.

In proposition 2, the same psychological constraints ψ are used in models (P, ψ) and $\left(P^C, \psi\right)$ that accommodate choices C. This leads to corollary 1 below.

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

Corollary 1 implies that Bob cannot conclude that Dee acted against her preferences in a binary choice unless binary choice is cyclic and cyclic preferences are ruled out. This holds for any permissibility assumptions (in section 5 we show that impermissibility assumptions also do not allow us to conclude that Dee acted against her preferences in a binary choice). Thus, the handicapped aversion claim *requires* cyclic choices *and* the assumption of preference orders. Yet, even these two assumptions are insufficient to uniquely identify handicapped aversion (the $S'-$model is also based on orders). Contextual information is also essential.

The handicap avoidance result does emerge under *order* rationalization theory with permissibility assumptions. Consider the assumption that Dee can rationalize watching the movie with the handicapped (e.g., she can rationalize z over y). In section 5.2, we show that this permissibility assumption and preference orders characterize the underlying conditions for the inference of handicapped avoidance from choices $\bar{C}$.

To sum up, the application of rationalization theory to the Synder et. al., study shows that the conclusion of handicapped avoidance does not follow directly from the evidence. In addition to observing cyclic binary choice it is necessary to assume that the choice of watching a movie with the handicapped person is psychologically permissible. Finally, despite observing cyclic choice, the decision maker must be assumed to have ordered preferences. Thus, the demonstration that Dee may have acted against her preferences ultimately rests on the same assumption commonly made in standard economics: agents have ordered preferences.

This example illustrates how rationalization theory can be used to characterize the conditions that sustain a given interpretation of the data. The example also shows that socially undesirable actions (e.g., avoiding the handicapped) are fostered by rationales that legitimize such behavior. Thus, legitimizing and delegitimizing rationales becomes a key to changing behavior that is entirely different from changing preferences or opportunities.

5. The Minimum Constraint Principle

Standard economic theory implicitly assumes Dee can rationalize all options. Similarly, psychology emphasizes the ease with which people rationalize. In this section we consider a version of rationalization theory that allows psychological constraints only as needed to accommodate choice.

Let (P, ψ) and (P', ψ') be two models that underlie a choice function C. The model (P, ψ) is *dominated* by (P', ψ') if P' is an order, and $\psi(B) \subseteq \psi'(B)$ for all issues $B \in \mathcal{B}$, with strict inclusion for some issue. So, if a model (P, ψ) is dominated then it uses more constraints than necessary to accommodate the observed choices. We refer to undominated models as those that satisfy the *minimum constraint principle*. Given a choice function C, let $\mathcal{P}^C \mathrm{x} \Psi^C$ be the *minimum constraint theory of rationalization*: the set of all models $(P, \psi) \in \mathcal{P} \mathrm{x} \Psi^{\mathcal{R}}$ that underlie C and are not dominated.

Definition 2. *Given a choice function C, a preference order P is identifiable if there exist permissibility assumptions $\mathcal{A}$ such that whenever x P y Bob infers that Dee prefers x to y with a theory of order $\mathcal{A}-$rationalization.*[4]

So, a preference is identifiable if, for some theory of behavior, Bob concludes that this preference is the only one that can accommodate the observed choices.

Theorem 1 Given choices C, a preference order P is identifiable if and only if there is a psychological constraint function ψ, such that the model (P, ψ) belongs to the minimum constraint theory of rationalization.

Theorem 1 fully characterizes preferences that are identifiable from data. The result is striking: a preference is identifiable from choice behavior and permissibility

[4] In the appendix we show (proposition $A.2$) that if choices are consistent with some model of rationalization then they are also consistent with an undominated model. Hence, the minimum constraint principle does not produce an empty set of models.

assumptions if and only if it satisfies the minimum constraint principle. In particular, if a preference does not satisfy the minimum constraint principle then it does not follow from the evidence, no matter which background assumptions are made over psychological constraints.

It is important to note that only permissibility assumptions are considered in Theorem 1. We now explain why impermissibility assumptions do not help identify preferences. An *impermissibility assumption* stipulates that Dee has a psychological constraint that prevents her form choosing an alternative. Formally, an impermissibility assumption $\mathcal{T}$ is a set of issues B_j and feasible options $y_j \in B_j$ such that Dee has no rationale that places y_j above all alternative options in B_j. We refer to both permissibility and impermissibility assumptions as contextual assumptions to highlight their dependence on the context. It might seem that impermissibility assumptions could help identify preferences. However, consider any $(y_j, B_j) \in \mathcal{T}$ and any choice function C. If $y_j = C(B_j)$ then Dee's choice contradicts the assumption that y_j is impermissible in B_j. On the other hand if $y_j \neq C(B_j)$ for every j then it is easy to see that given any model (P, ψ) that underlies C there exists an alternative model (P, ψ') that also underlies C such that ψ' satisfies $\mathcal{T}$ (i.e., $\psi'(B)$ comprise of all options in $\psi(B)$ minus those that are assumed by $\mathcal{T}$ to be impermissible). Thus, impermissiblity assumptions cannot aid in the identification of preferences from a choice function. Only permissibility assumptions can help identify preferences. So, a valid interpretation of the data can be expressed by an undominated model and each undominated model can be differentiated by the permissibility assumptions that support it.

The intuition in Theorem 1 is that if a model (P, ψ) is dominated then there is an alternative model (P', ψ') with fewer constraints that also accommodates the observed choices. So, any permissibility assumption that is satisfied by ψ is also satisfied by ψ'. Thus, (P, ψ) cannot be the only way to accommodate choice under any set of contextual assumptions.

We now show that the minimum constraint theory of rationalization has several desirable properties. First, the only undominated model that can underlie non-anomalous behavior is the one without any psychological constraints. So,

The Comparability Claim If choices are non-anomalous then standard theory and the minimum constraint theory of rationalization reveal identical preferences.

The comparability claim shows that *the minimum constraint theory of rationalization does not contradict, or even modify, standard economics.* Standard economic theory applies when no violations of WARP have been observed. In this case, minimum constraint rationalization theory generates identical preferences as standard economic theory. Therefore, while rationalization theory can be used to understand anomalous behavior, it does not impair the insights of standard economics in the analysis of non-anomalous behavior.

The psychological idea of rationalization can be seen not as a replacement for standard theory but as a broadening of standard economics that allows for behavior patterns previously regarded as anomalous. To make this claim formal, let's say that a choice function is *acyclic* if the binary choices do not form a cycle.

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 in a binary choice) then Bob infers that Dee prefers x to y by minimum constraint rationalization theory.*

Proposition 3 shows that when a choice function is acyclic then, among all possible preferences (orders or not), the only surviving preference relation is the order defined by the binary choices. Thus, a complete identification of preferences is now broadened to include behavior patterns in which choices are 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 behavior is anomalous, binary choices fully reveal preferences. The intuition is as follows: consider a model (P, ψ) that underlies the choice function C. Then, by proposition 2, $\left(P^C, \psi\right)$ also underlies C. Hence, $\left(P^C, \psi'\right)$ also underlies C, where ψ' is unconstrained in binary choices and equals ψ elsewhere. Then, an undominated model must be unconstrained in binary choices.

We conclude this section by showing that, for any preferences that are part of an undominated model, psychological constraints are uniquely identified.

Proposition 4. *If two undominated models, (P, ψ) and (P, ψ'), underlie choices C then the psychological constraints must coincide, (i.e., $\psi = \psi'$).*

By proposition 4, if Dee's preferences are revealed under the minimum constraint rationalization theory then Dee's constraints are also fully identified. The intuition is that the options that Dee prefers and does not choose are revealed to be impermissible.

Propositions 3 and 4 allow a complete inference of preferences and constraints when observed binary choices are acyclic. As in standard theory, these inferences depend only on choice and are situation-specific.

5.1. Difficult Choice Anomaly

Consider the following choice function: $C(e_1, e_2) = e_1$; $C(e_1, n) = e_1$; $C(e_2, n) = e_2$; and $C(e_1, e_2, n) = n$. This pattern is anomalous because n is rejected over e_1 and also over e_2 separately, but n is chosen over both e_1 and e_2 when they are simultaneously available. From this pattern of choice alone, we can determine preferences and constraints. By proposition 3, minimum constraint rationalization theory reveals that $e_1 \ P \ e_2 \ P \ n$.[5] Proposition 4 reveals that the only binding psychological constraint occurs in the issue $\{e_1, e_2, n\}$ in which neither e_1 or e_2 is psychologically feasible.

The choices above are consistent with anecdote about Thomas Schelling (as told by Shafir et. al., (1993)) who, one an occasion, had decided to buy an encyclopedia. Upon arriving at the bookstore, had only one encyclopedia been available (e_1 or e_2), he would have happily bought it. However, he was presented with two encyclopedias. Finding it difficult to choose he ended up buying neither (n).

The pattern of behavior above is an acyclic behavioral anomaly often called a *difficult choice.* Tversky and Shafir (1992), Simonson and Tversky (1993), among others, noted the difficult choice anomaly in several experiments. In a field experiment, Iyengar and Lepper (2000) observed 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 tells us that the difficult choice behavior is not necessarily the result of incomplete or unordered preferences, but can result from the inability to rationalize choice when many alternatives are provided. More generally, introducing new alternatives may produce new psychological constraints, while reducing available options may relax them.

5.2. Cycles

The minimum constraint rationalization theory provides an elegant extension of standard economics when choice is acyclic. When choice is cyclic it is often possible to find

[5] In fact, the inference that $e1 \ P \ n$ and $e2 \ P \ n$ follows from Proposition 1 and does not require the minimum constraint principle.

a unique preference order that produces the observed behavior even without recourse to the minimum constraint principle. These examples are available upon request. So, in many cases where choices are anomalous, the structure of 3.1 that rationalization imposes on psychological constraints is sufficient to produce a full identification of preference orders and constraints. As shown by proposition 1, the minimum constraint principle is essential in the identification of preferences when choice is not anomalous. Yet, the addition of the minimum constraint principle may reveal a unique preference order in some cycles as well (examples also available upon request). However, there are behavioral patterns in which multiple undominated models can underlie observed choices. In such cases, there are several valid interpretations of the data. Then, contextual information may help resolve the remaining ambiguity over preferences and constraints.

Consider the three-alternative cycle: x over y, y over z and z over x (let's say y chosen when all three alternatives are available). This is the behavioral pattern $\bar{C}$ in the handicapped aversion example. Both the $S-$model and the $S'-$model are undominated and can accommodate $\bar{C}$. So, Bob may still want to differentiate between these models. As shown above, impermissibility assumptions won't help. However, as shown by Theorem 1, each undominated model can be characterized by permissibility assumptions. So, Bob may be able to differentiate between models on the basis of such assumptions. We now illustrate this method.

Consider the contextual assumption that y is rationalizable in (y, z), (i.e., $y \in \psi(y, z)$). In the example, this means that Dee can rationalize watching a movie with the handicapped when the alternative is to see another movie alone. Then, from y chosen over z it follows that Dee prefers y to z. By proposition 1, Dee prefers x to y. So, Dee's revealed preference order is model S; she prefers x to y to z: she prefers to avoid the handicapped. Now consider the contextual assumption that x is rationalizable in (x, z), (i.e., $x \in \psi(x, z)$). That is, Dee can rationalize watching a movie alone when the alternative is to see the same movie with the handicapped. Then, from z chosen over x it follows that Dee prefers z to x. By proposition 1, Dee prefers x to y. So, Dee's revealed preference order is model S'; she prefers z to x to y: she does not prefer to avoid the handicapped. Bob's view on handicapped avoidance must depend on his judgement over these two critical permissibility assumptions.

While rationalization theory extends standard theory in a natural way it is appropriate to ask about the empirical scope of rationalization theory. When does

the theory apply and when should it be rejected? We begin by providing a basic characterization result.

6. Testing rationalization theory

In this section, we characterize the choice functions that are consistent with basic rationalization theory. In the appendix, we also show a characterization of the empirical content of the theory of $\mathcal{A}-$rationalization and the theory of order $\mathcal{A}-$rationalization.

Weak WARP A choice function C satisfies the Weak WARP iff

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

Weak WARP is a familiar and natural relaxation of WARP (see Manzini and Mariotti (2007, 2010)). A violation of Weak WARP occurs if the following two violations of WARP are observed. First, x is chosen over y and y is chosen in a larger issue B_1 that contains x. By proposition 1, under rationalization theory, Dee prefers x over y. In the second violation, y is chosen in B_1 and x is chosen in an even larger issue B_2 that contains y. By proposition 1, Dee prefers y over x. This contradiction shows that behavioral patterns consistent with basic rationalization theory must satisfy Weak WARP. The converse also holds.

Proposition 5. *A choice function C is consistent with basic rationalization theory if and only if it satisfies Weak WARP.*

So, rationalization theory can accommodate all behavioral anomalies that satisfy Weak WARP (such as cycles) and is consistent with any observed choices over three alternatives. Moreover, the logical structure imposed on psychological constraints by rationalization theory guarantees that the theory is testable even when the constraints themselves are unobservable.

7. Rationalization Theory: Applications

7.1. Interpreting (Perhaps Contradictory) Speech

Assume Dee tells Bob that it would be acceptable for someone like her to choose x over y. Bob may interpret Dee's statement as evidence that she can rationalize x

over y. It does not mean that Dee will choose x over y, but it does mean that if Dee chooses y then she must prefer it to x. 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. Consider an accountant who says that she could never do anything illegal, but could understand why someone would. 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.

7.2. Acceptable Rationales and Social Norms

The economics literature emphasizes that behavior can change in response to changes in opportunities and incentives. Rationalization theory provides an additional mechanism for behavioral change due to a change in acceptable rationales. Consider the well-known experiment of Gneezy and Rustichini (2000) in which they introduced small fines for lateness at Israeli day care centers. They found that, rather than reduce lateness, the fines increased lateness. Moreover, the lateness persisted even when the fines were removed. This behavior is consistent with rationalization theory under the assumption that the introduction of a fine made greater levels of lateness rationalizable and that the removal of the fine did not eliminate the new rationale.

In laboratory experiments, Ariely and Mazar (2006) found that a willingness to cheat depends upon whether subjects thought that 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. This suggests that social norms are important drivers of behavior.

Suppose that the only acceptable rationales for Dee are those that are socially acceptable. If Dee is informed that most guests reuse their towels then she learns that most people don't have a rationale for not reusing their towels. She infers there must not be socially acceptable rationale for not reusing towels at hotels and so she reuses her towels. Thus, a descriptive norm eliminates acceptable rationales and imposes a psychological constraint.

Relaxing psychological constraints may also have an impact on behavior. Marks (2005) reports the following story about the introduction in the 1940s of powdered cake mix into consumer markets. The new cake mix allowed consumers to simply add water and bake. The product saved time but the company, General Mills, was

surprised by the poor sales. Research by psychologists Burleigh Gardner and Ernest Dichter at the time suggested that powdered eggs should be left out so that fresh eggs would have to be added by consumers. Adding eggs made the mix marginally less convenient, but Gardner and Ernest believed that consumers who bought cake mix were psychologically invested in the idea that their cake should be home made. Requiring consumers to add eggs allowed consumers to rationalize calling their cake home made and so sales increased.[6]

7.3. Game Theory

Consider the sequential prisoner dilemma game in Figure 1 below.

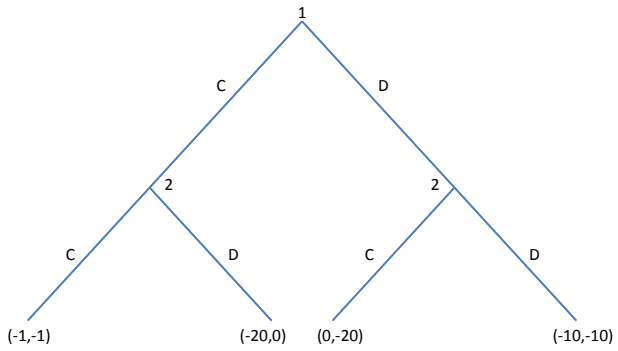

Figure 1. Prisoner's Dilemma

Suppose the players prefer the largest payoff possible but can only rationalize actions that are socially acceptable: cooperation can always be rationalized but defection can only be rationalized if the other has defected or is expected to defect following cooperation. So, player 2 cooperates following 1's cooperation and defects following player 1's defection. Similarly, player 1 cooperates as well. In effect rationalization theory produces behavior that resembles reciprocity. See Rabin (1993), Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) for models of reciprocity

[6]Finding 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

and Spiegler (2002) and (2004) for game-theoretic models where players must justify their chosen actions.

Rationalization theory is a constrained optimization process of a single and well-defined objective. Thus, it is straightforward to integrate it within game theory if preferences and psychological constraints are common knowledge. However, a general introduction of rationalization into game theory would require extending the theory to allow for choice over lotteries and uncertainty over payoffs and rationales. This is beyond the scope of this paper.

7.4. Legal Studies

Assume that Dee maximizes her preferences but is psychologically constrained to obey the law. Bob may then observe cyclic choices (see Katz and Sandroni (2010) for examples in branches of law such as self-defense, duress, necessity, and negligence). Psychological constraints may be particularly relevant for understanding the behavior of legislators, bureaucrats and judges. The key point is that one should expect a divergence between traditional criteria of rationality (e.g., transitivity) and the behavior of rational, but psychologically constrained, agents.

7.5. Differentiating Theories

Many behavioral theories can accommodate anomalies. Thus, it is natural to ask how to differentiate between them. Consider the marketing field study of Berger and Smith (1997). They observe that some donors (to universities) choose to make a small solicited contribution (s) over no contribution (n), but if donors are solicited to make either a small or a large contribution (l) then they choose not to contribute. These two choices, $C(s,n) = s$ and $C(s,n,l) = n$, are anomalous. Depending on what is chosen between n and l and also s and l, we may end up with a cycle or an anomaly known in the literature as the *attraction effect*. Both patterns can be accommodated by rationalization theory. However, regardless of what the two unobserved choices might be, by proposition 1, Bob must infer that Dee prefers a small contribution over no contribution. Consider the contextual assumption that Dee *can* rationalize a small donation. This contradicts her choice of no donation. So, under this contextual assumption, rationalization should not be considered a viable explanation for this phenomena. Thus, contextual assumptions not only help select among

alternative models of rationalization but may also help circumscribe the application of the theory itself. We refer the reader to Ok et al (2008), Clippel and Eliaz (2009), and Cherepanov, Feddersen and Sandroni (2010) for theories that accommodate the attraction effect. See also Masatlioglu and Nakajima (2007), Masatlioglu and Ok (2007), Eliaz and Ok (2006) for related models.

7.6. Welfare

Welfare analysis is an important topic when agents face psychological constraints. Suppose that Dee wants to have a life-saving medical procedure but would choose against it because of a moral prohibition against such procedures. Should someone acting on her behalf choose according to Dee's revealed preferences or according to the choice she would make subject to her psychological constraints. A variety of perspectives have been offered on this issue (see, for example, Mill (1860), Thaler and Sunstein (2003)). Rationalization theory can reveal Dee's preference and constraints and, hence, determine when they clash, but the welfare implications of such clashes are still unresolved.

8. Conclusion

We develop a formal model of the psychological idea of rationalization. Our starting point is that the inability to rationalize may place unobservable psychological constraints on choice. Rationalization theory imposes logical structure on psychological constraints and, thereby, guarantees that the theory is testable. Under minimum constraint rationalization theory preferences and constraints are uniquely revealed across a wide variety of choice patterns. When observed choice is not anomalous, minimum constraint theory reveals the same preferences as standard economics. When binary choice behavior is anomalous but acyclic, unique preferences and constraints are revealed from choice alone. When ambiguity over preferences remains, evidence that behavior is permissible may be used to reduce ambiguity and to reject the model outright. By combining the psychological idea of rationalization with the economic idea of ordered preferences and constrained choice we get a new theory that can deepen and extend analysis in both disciplines.

While we consider a decision theoretic framework, the model can serve as a foundation for novel strategic analyses. In particular, rationalization provides a testable

formal structure that can help understand why debates about seemingly abstract principles might become such a central feature of social life: such debates can change behavior without changing preferences or the feasibility of choice.

9. Appendix : Proofs and Extended Results

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

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 $\mathcal{A} = \{(y_i, B_i);\ y_i \in B_i\ i = 1, ..., n\}$, let $P^{C,\mathcal{A}}$ be the binary relation such that $x\ P^{C,\mathcal{A}}\ y$ if and only if

$x = C(B)$ and $y = C(B^*)$ for some pair (B, B^*) of nested issues that violates WARP or

$x = C(B)$ and for some $(y_i, B_i) \in \mathcal{A},\ y = y_i \in B \subseteq B_i$.

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

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

By definition,

$$C(B)\ P^{C,\mathcal{A}}\ y \text{ for any } y \in \psi^{C,\mathcal{A}}(B),\ y \neq C(B) \tag{9.1}$$

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

$$\psi^{C,\mathcal{A}}(\tilde{B}) \bigcap B \subseteq \psi^{C,\mathcal{A}}(B)$$

This follows because if $z \in B$ and $z \in \psi^{C,\mathcal{A}}(\tilde{B})$ then we can assume, without loss of generality, that $z \neq C(B)$. Otherwise $z = C(B)$ and so, $z \in \psi^{C,\mathcal{A}}(B)$. 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, $z \in \psi^{C,\mathcal{A}}(B)$. Hence, it follows from $z \in \psi^{C,\mathcal{A}}(\tilde{B})$ 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, $z \in \psi^{C,\mathcal{A}}(B)$. So, $\psi^{C,\mathcal{A}} \in \Psi^{\mathcal{A}}$.

Lemma 1. *If $(P, \psi) \in \mathcal{P}\mathrm{x}\Psi^{\mathcal{A}}$ underlies C then $x\ P^{C,\mathcal{A}}\ y \Longrightarrow x\ P\ y$*

Proof : Assume that $x = C(B)$ and $y = C(B^*)$ for some pair (B, B^*) of nested issues that violates WARP. Then, $y \in \psi(B^*)$ (because $y = C(B^*)$) and $y \in B$ (because $C(B^*) \in B$). So, by 3.1, $y \in \psi(B)$. Hence, $x\ P\ y$ (because (P, ψ) 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, $y_i \in \psi(B_i)$ (because $\psi \in \Psi^{\mathcal{A}}$). By 3.1, $y_i \in \psi(B)$. Hence, $x\ P\ y = y_i$.■

Proposition A.1 A choice function C consistent with $\mathcal{A}-$rationalization theory $\mathcal{P}\mathrm{x}\Psi^{\mathcal{A}}$ if and only if $P^{C,\mathcal{A}}$ is asymmetric. A choice function C is consistent with $\mathcal{A}-$rationalization order theory $\mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$ if and only if $P^{C,\mathcal{A}}$ is acyclic.

Proof : Assume that a choice function C is consistent with $\mathcal{A}-$rationalization theory $\mathcal{P}\mathrm{x}\Psi^{\mathcal{A}}$. Let $(P, \psi) \in \mathcal{P}\mathrm{x}\Psi^{\mathcal{A}}$ be a model that underlies C. Assume, by contradiction, that $P^{C,\mathcal{A}}$ is not asymmetric. Then, for some $x \neq y$, $x\ P^{C,\mathcal{A}}\ y$ and $y\ P^{C,\mathcal{A}}\ x$. By Lemma 1, $x\ P\ y$ and $y\ P\ x$. This contradicts, $P \in \mathcal{P}$. Now assume that $P^{C,\mathcal{A}}$ is asymmetric. Then, by 9.1, $(P^{C,\mathcal{A}}, \psi^{C,\mathcal{A}}) \in \mathcal{P}\mathrm{x}\Psi^{\mathcal{A}}$ underlies C.

Next assume that a choice function C is consistent with order $\mathcal{A}-$rationalization theory $\mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$. Let $(P, \psi) \in \mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$ be a model that underlies C. Assume, by contradiction, that $P^{C,\mathcal{A}}$ is cyclic. Then, Lemma 1, P is also cyclic. This contradicts, $P \in \mathcal{P}^o$. Now assume that $P^{C,\mathcal{A}}$ is acyclic. By topological ordering, $P^{C,\mathcal{A}}$ may be extended (not necessarily uniquely) to an order (see Cormen et al. (2001, pp.549–552)). Let $\bar{P}$ be an arbitrary order that extends $P^{C,\mathcal{A}}$. Then, by 9.1, $(\bar{P}, \psi^{C,\mathcal{A}}) \in \mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$ underlies C.■

Proposition $A.1$ characterizes the empirical content of (order) $\mathcal{A}-$rationalization theory.

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

If C is consistent with rationalization theory $\mathcal{P}\mathrm{x}\Psi^{\mathcal{R}}$then, by Proposition $A.1$, $P^{C,\varnothing}$ is asymmetric. Hence, $\left(P^{C,\varnothing}, \psi^{C,\varnothing}\right) \in \mathcal{P}\mathrm{x}\Psi^{\mathcal{R}}$ 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 $x\ P^{C,\varnothing}\ y$. Thus, consider the binary relation $\bar{P}$ such that $y\ \bar{P}\ x$ and for all other pairs of alternatives $\bar{P}$ is identical to $P^{C,\varnothing}$. Then, $\bar{P}$ is still asymmetric and $\left(\bar{P}, \psi^{C,\varnothing}\right) \in \mathcal{P}\mathrm{x}\Psi^{\mathcal{R}}$ still underlies C.■

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

Proposition A.2 A choice function that is C is consistent with basic rationalization theory is consistent with the minimum constraint theory of rationalization.

Proof : Let's define the partial order $\geqq$ on psychological constraint function such that $\psi' \geqq \psi$ if and only if $\psi(B) \subseteq \psi'(B)$ for all issues $B \in \mathcal{B}$. Given that the set of all alternatives A is finite, there is an $\geqq-$maximal element, ψ^*, in the set of $\{\psi :$ for some P, (P,ψ) underlies $C\}$. So, for some P^*, (P^*, ψ^*) is an undominated model that underlies C.■

Proof of Theorem 1 : Assume that an order P is identifiable. Let $\overline{\psi}$ be the $\geqq-$maximal element in $\{\psi :$ for some P, (P,ψ) underlies $C\}$. So, (P,ψ) underlies C. Assume, by contradiction, that $\left(P, \overline{\psi}\right)$ is dominated. Then, there exists a model (P', ψ') that also underlies C such that $\psi(B) \subseteq \psi'(B)$ for all issues $B \in \mathcal{B}$, with strict inclusion for some issue $B \in \mathcal{B}$. Then, $P' \neq P$ ($P' = P$ contradicts the $\geqq-$maximality of $\overline{\psi}$). Consider the model $\left(P', \overline{\psi}\right)$. First, $\left(P', \overline{\psi}\right)$ also underlies C because for any issue B, $C(B)\ P'\ y$ for all $y \in \psi'(B) \supseteq \psi(B)$. So, $C(B)\ P'\ y$ for all $y \in \overline{\psi}(B)$. Moreover, for any series of permissibility assumptions $\mathcal{A}$, if $\left(P, \overline{\psi}\right) \in \mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$ then $\left(P', \overline{\psi}\right) \in \mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$. Thus, P is identified by $\mathcal{A}$.

Assume that P is an order and (P,ψ) belongs to the minimum constraint theory of rationalization. Let $\mathcal{A}$ be the series of permissibility assumptions defined $(y,B)\in\mathcal{A}$ if and only if $y\in\psi(B)$. Assume, by contradiction, that P is not identified by the theory of order $\mathcal{A}-$rationalization. Then, there exists a model $(P',\psi')\in\mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$ that underlies C, with $P'\neq P$. Now, $\psi'\geqq\psi$ (because $\psi'\in\Psi^{\mathcal{A}}$). So, $\psi'=\psi$ (otherwise (P,ψ) is a dominated model). Thus, (P',ψ) underlies C. Now let a and b be two alternatives such that $a\ P\ b$ and $b\ P'\ a$ (these alternatives exist because $P'\neq P$). Let $\hat{\psi}$ be identical to ψ on all issues $B\neq\{a,b\}$ and $\hat{\psi}\{a,b\}=\{a,b\}$. By definition, if $\{a,b\}\subseteq\tilde{B}$ then $\hat{\psi}(\tilde{B})\bigcap\{a,b\}\subseteq\hat{\psi}\{a,b\}$ and $y\in\hat{\psi}(B)$ if $(y,B)\in\mathcal{A}$. Thus, $\hat{\psi}\in\Psi^{\mathcal{A}}$. Now either $b\in\psi\{a,b\}$ or $b\notin\psi\{a,b\}$. In the later case, $\psi\{a,b\}=\{a\}$, $C\{a,b\}=a$ and $(P,\hat{\psi})$ underlies C (because $a\ P\ b$). Thus, (P,ψ) is a dominated model. A contradiction. In the former case $b\in\psi\{a,b\}$. Then, $C\{a,b\}=b$ (because $b\ P'\ a$ and (P',ψ) underlies C). Thus, $\left(P',\hat{\psi}\right)$ underlies C. In addition, $\psi\{a,b\}=\{b\}$ ($\psi\{a,b\}=\{a,b\}$ would contradict $a\ P\ b$ and (P,ψ) underlies C). Hence, (P,ψ) is a dominated model. A contradiction.■

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

Let $B\neq\{x,y\}$ be an issue. Let $z\in\psi'(B)=\psi(B)$, $z\neq C(B)$. Note that $\{C(B),z\}\subseteq B$ and $C(B)\in\psi(B)=\psi'(B)$. So, $\{C(B),z\}\subseteq B\bigcap\psi'(B)$ and $\{C(B),z\}\subseteq B\bigcap\psi(B)$. Hence, $\psi(\{C(B),z\}=\psi'(\{C(B),z\})=\{C(B),z\}$. It follows that $C(B)\ P\ z$ (because $z\in\psi(\{C(B),z\}$ and (R,ψ) underlies C). Hence, $C\left(\{C(B),z\}\right)=C(B)$ (because (R,ψ) underlies C). By definition, $C(B)\ P^C\ z$. Moreover, $C(\{x,y\})=x$ (because $\psi\{x,y\}=\{x\}$). So, $x\ P^C\ y$.Hence, $(P^C,\psi')\in\mathcal{P}^o\mathrm{x}\Psi^{\mathcal{R}}$ underlies C and (P,ψ) is dominated by (P^C,ψ'). Thus, for every pair of alternatives $\{x,y\}$, $\psi\{x,y\}=\{x,y\}$. Given that (P,ψ) underlies C it now follows that $P=P^C$.
■

Proof of Proposition 4 : Let (P,ψ) and (P,ψ') be two undominated models that underlie choices C. Assume, by contradiction, that some some issue $\bar{B}$, $\psi(\bar{B})\neq\psi'(\bar{B})$. Let $\hat{\psi}$ be defined by $\hat{\psi}(B)=\psi(B)\bigcup\psi'(B)$. By definition, either ψ or ψ' (or both)

imposes more constraints than $\hat{\psi}$. Now, $\left(P, \hat{\psi}\right)$ underlies C because $C(B)\ P\ y$ for every $y \in \psi(B)$ and for every $y \in \psi'(B)$. In addition, if $B \subseteq \tilde{B}$ then $\hat{\psi}(\tilde{B}) \bigcap B \subseteq \hat{\psi}(B)$. This follows because $\psi(\tilde{B}) \bigcap B \subseteq \psi(B)$ and $\psi'(\tilde{B}) \bigcap B \subseteq \psi'(B)$. So, $\hat{\psi} \in \Psi^{\mathcal{R}}$. Thus, either (P, ψ) or (P, ψ'), or both, are dominated models. A contradiction.■

Proposition A.3 Consider a choice function C consistent with $\mathcal{A}-$rationalization theory $\mathcal{P}\mathrm{x}\Psi^{\mathcal{A}}$. 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 $(y_i, B_i) \in \mathcal{A}$, $y = y_i \in B \subseteq B_i$.

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

Proposition A.4 Consider a choice function C consistent with order $\mathcal{A}-$rationalization theory $\mathcal{P}^o\mathrm{x}\Psi^{\mathcal{A}}$. Then, Bob infers that Dee prefers z_1 to z_k if there exists a chain z_{i+1}, $i = 0, ..., k-1$, such that Bob infers that Dee prefers z_i to z_{i+1} by $\mathcal{A}-$rationalization theory.

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

Proposition $A.3$ and $A.4$ fully characterizes the preference inferences that can made under (order) $\mathcal{A}-$rationalization theory. Take the basic rationalization theory as a benchmark. Proposition $A.3$ shows that the *only* additional inferences over preference that follows from permissibility 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 $A.4$ 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.

We now return to the basic rationalization theory (i.e., $\mathcal{A} = \varnothing$ 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 $C(B_1^*) = C(B_2)$.

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

By proposition $A.1.$, this axiom fully demarcates the choice functions that can and cannot be accommodated by the basic rationalization theory because two pairs of nested issues that violate WARP with reversed choices if and only if $P^{C,\varnothing}$ is asymmetric

Proposition A.5 The irreversibility axiom holds if and only if Weak WARP holds.

Assume that Weak WARP 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, Weak WARP 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, Weak WARP does not hold.■

The proof of proposition 5 is a direct corollary of Propositions $A.1$ and $A.5$.

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] Ariely, D. and N. Mazar (2006) "Dishonesty in Everyday Life and its Policy Implications" *Journal of Public Policy and Marketing*, 25-1, 117-126.

[5] Aronson, E. and A. Pratkanis (2001) "Age of Propaganda: The everyday use and abuse of presuasion," Henry Holt and Company. Revised Edition.

[6] Bénabou, R. and J. Tirole (2002) "Self-Confidence and Personal Motivation," *The Quarterly Journal of Economics*, 117 (3), 871–915.

[7] 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.

[8] Bernheim, D. (1984) "Rationalizable Strategic Behavior," *Econometrica*, 52 (4), 1007-1028 .

[9] Bernheim, D. and A. Rangel (2009) "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *The Quarterly Journal of Economics,* 124 (1), 51-104.

[10] Bolton, G. and A. Ockenfels (2000) "ERC A Theory of Equity, Reciprocity and Competition," *American Economic Review,* 90 (1), 166-193.

[11] Carillo, J. and T. Mariotti (2000) "Strategic Ignorance as a Self-Discipline Device," *Review of Economic Studies*, 67 (3), 529-544.

[12] Chambers. C. and T. Hayashi (2008) "Choice and Individual Welfare," mimeo, Caltech.

[13] Chen, K. (2008) "Rationalization and Cognitive Dissonance: Do Choices Affect or Reflect Preferences," mimeo.

[14] Cherepanov, V., Feddersen, T, and A. Sandroni (2010) "Revealed Preferences and Aspirations in Warm Glow Theory." Working paper, Northwestern University.

[15] Cormen, T., C. Leiserson, R. Rivest, and C., Stein (2001) "Introduction to Algorithms," MIT Press and McGraw-Hill. Second Edition.

[16] Clippel, G. and K. Eliaz (2009) "Reason-Based Choice: A Bargaining Rationale for the Attraction and Compromise Effects," Working paper, Brown University.

[17] Dietrich, F. and C. List (2010) "A Reason-based Theory of Rational Choice," mimeo.

[18] Fehr, E., and K. Schmidt (1999) "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114, 817–68.

[19] Fudenberg, D. and D. Levine (2006) "A Dual-Self Model of Impulse Control," *American Economic Review*, 96 (5), 1449-1476.

[20] Gneezy, U., and A., Rustichini (2000) "A fine is a price," *The Journal of Legal Studies*, Vol 29, No. 1 (January), pp1-17.

[21] Goldstein, N., and Cialdini, R. (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.

[22] Green, J., and D. Hojman (2007) "Choice, Rationality and Welfare Measurement," Harvard University, Working Paper.

[23] Gul, F. and W. Pesendorfer (2005) "The Revealed Preference Theory of Changing Tastes," *Review of Economic Studies* 72 (2) , 429–448.

[24] Heller, Y. (2009) "Justifiable Choice," mimeo, Tel Aviv University.

[25] 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.

[26] Jones, E. (1908) "Rationalisation in Every-day Life," *Journal of Abnormal Psychology*, 161-169.

[27] Kalai, G., A. Rubinstein, and R. Spiegler (2002) "Rationalizing Choice Functions by Multiple Rationales," *Econometrica*, 70 (6), 2481-2488.

[28] Katz, L. and A. Sandroni (2010) "Why Law Breeds Cycles," mimeo, University of Pennsylvania.

[29] Lehrer, E. and R. Teper (2009) "Justifiable Preferences," mimeo, Tel Aviv University.

[30] Manzini, P. and M. Mariotti (2007) "Sequentially Rationalizable Choice" *American Economic Review* **97**-5, 1824-1839.

[31] Manzini, P. and M. Mariotti. (2010) "Categorize then Choose: Bondedly Rational Choice and Welfare," forthcoming *Journal of the European Economic Association.*

[32] Masatlioglu, Y. and D. Nakajima (2007) "A Theory of Choice by Elimination," mimeo.

[33] Masatlioglu, Y. and E. Ok (2007) "Status Quo Bias and Reference-Dependent Procedural Decision Making," mimeo.

[34] Mill J.S. (1860): *On Liberty,* P.F. Collier & Son, London.

[35] Moulin, H. (1985) "Choice Functions over a Finite Set: A Summary," *Social Choice and Welfare*, 2, 147-160.

[36] Ok, E., P. Ortoleva, and G. Riella (2008) "Rational Choice with Endogenous Reference Points," mimeo.

[37] Pearce D. (1984) "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52, (4), 1029-1050.

[38] Rabin, M. (1993) "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83 (4), 1281–1302.

[39] Rabin, M. (1995) "Moral Preferences, Moral Constraints, and Self-Serving Biases," mimeo.

[40] Roth, A. (2007) "Repugnance as Constraints on Markets," mimeo Harvard University.

[41] Salant, Y. and A. Rubinstein (2006) "Two Comments on the Principle of Revealed Preferences," mimeo.

[42] Salant, Y. and A. Rubinstein (2006a) "A Model of Choice from Lists," *Theoretical Economics*, 1, 3-17.

[43] Samuelson, P. (1938) "The Empirical Implications of Utility Analysis," *Econometrica*, 6 (4), 344-356.

[44] Shafir, E., I. Simonson, and A. Tversky (1993) "Reason-based Choice," *Cognition* 49, 11-36.

[45] Simonson, I and A. Tversky (1993) "Context-Dependent Preferences," *Management Science*, 39 (10), 1179-1189.

[46] Sprumont, Y. (2000) "On the Testable Implications of Collective Choice Theories," *Journal of Economic Theory*, 93, 205-232.

[47] Spiegler, R. (2002) "Equilibrium in Justifiable Strategies: A Model of Reason-Based Choice in Extensive-Form Games" (2002), *Review of Economic Studies* 69, 691-706.

[48] Spiegler, R. (2004) "Simplicity of Beliefs and Delay Tactics in a Concession Game," *Games and Economic Behavior* 47 (1), 200-220.

[49] 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.

[50] Thaler R. and C. Sunstein (2003): "Libertarian Paternalism", *American Economic Review* 93: 175-179.

[51] Tversky, A. and E. Shafir (1992) "Choice Under Conflict: The Dynamics of Deferred Decision," *Psychological Science*, 3 (6), 358-361.

[52] 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.