INTERDEPENDENT PREFERENCES AND STRATEGIC DISTINGUISHABILITY

By

Dirk Bergemann, Stephen Morris and Satoru Takahashi

September 2010
Revised February 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1772R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Interdependent Preferences and Strategic Distinguishability*

Dirk Bergemann† Stephen Morris‡ Satoru Takahashi§

February 2011

Abstract

A universal type space of interdependent expected utility preference types is constructed from higher-order preference hierarchies describing (i) an agent's (unconditional) preferences over a lottery space; (ii) the agent's preference over Anscombe-Aumann acts conditional on the unconditional preferences; and so on.

Two types are said to be strategically indistinguishable if they have an equilibrium action in common in any mechanism that they play. We show that two types are strategically indistinguishable if and only if they have the same preference hierarchy. We examine how this result extends to alternative solution concepts and strategic relations between types.

KEYWORDS: Interdependent Preferences, Higher-Order Preference Hierarchy, Universal Type Space, Strategic Distinguishability.

JEL CLASSIFICATION: C79, D82, D83.

*The first two authors acknowledge financial support through NSF Grant SES 0851200. We are grateful for comments from seminar/conference participants at Columbia, Chicago, Harvard/MIT, HEC, Kyoto, Northwestern, NYU, Oxford, Penn, SAET, Yale, Warwick and the Econometric Society World Congress in Shanghai.

†Yale University, dirk.bergemann@yale.edu

‡Princeton University, smorris@princeton.edu

§Princeton University, satorut@princeton.edu

1 Introduction

Economists often assume that agents' preferences are interdependent for informational or psychological reasons. We know how to use Harsanyi type spaces to represent many kinds of such interdependence of preferences. In this paper, we characterize when two types are strategically distinguishable in the sense that they are guaranteed to behave differently in some finite mechanism mapping actions to outcomes.

Our characterization uses a universal type space of interdependent, higher-order, preferences of a finite set of agents, analogous to the universal space of higher-order beliefs introduced by Mertens and Zamir (1985). We assume common certainty that (i) agents are expected utility maximizers; (ii) agents are not indifferent between all outcomes; and (iii) there is a worst outcome for each agent. The universal space is mathematically isomorphic to the Mertens-Zamir universal belief space (although it has a very different interpretation). We show that two types are strategically distinguishable if and only if they map to different points in the universal space of interdependent preferences.

This result gives a clean and straightforward answer to the question: what can you observe (and be certain to observe) about agents' interdependent preferences by seeing how they play games, i.e., behave in strategic environments? Our answer is:

1. You can learn an agent's first order (or unconditional) preferences: what are his preferences over outcomes unconditional on anything other agents do or say?

2. Since you can learn all agents' unconditional preferences, you can also learn an agent's second order preferences: what are his preferences over acts that are contingent on the first order preferences of other agents?

3. And then you can learn his third order preferences. And so on.

You cannot learn any more than this. This implies, in particular, that it is not possible to distinguish between informational and psychological reasons for interdependence. And it implies that interdependence of preferences can be observed only when there is uncertainty about preferences, i.e., when I expect my preference to change upon observing your preferences.

There are (at least) a couple of reasons why we believe that a systematic study of strategic distinguishability may be of interest. First, economists' traditional view of preferences is that they are not directly observed but are best understood as being revealed by agents' choices in actual or hypothetical decision problems, and there is a developed revealed preference theory of individual

choice behavior; we see this paper as being a step towards a strategic revealed preference theory.[1] Second, the content of the specific modelling assumptions is not always transparent and this is especially true when talking about interdependent preferences. By mapping all types into a canonical universal interdependent type space, we provide a clear operational definition of interdependent types.

Our main result concerns one solution concept, equilibrium, and one equivalence class on agents' interdependent types, strategic indistinguishability. We also discuss what happens if we consider an appropriate but very permissive definition of rationalizability for our environment—dubbed interim preference correlated rationalizability (IPCR)—and an alternative, more refined, equivalence class on agents' types: two types are said to be strategically equivalent if they have the same set of rationalizable actions in all strategic environments (strategic distinguishability only required a non-empty intersection of those sets). We show that the same universal interdependent preference space characterizes strategic distinguishability for IPCR, and thus for any solution concept which refines IPCR and coarsens equilibrium. We also show that the universal interdependent preference space characterizes strategic equivalence for IPCR, so that, for IPCR, two types are strategically distinguishable if and only if they are strategically equivalent. But for equilibrium, more information than that contained in the universal interdependent preference space is required to capture strategic equivalence (as shown by an example in Section 3).

We maintain the worst outcome assumption in order to exclude trivial types that are completely indifferent over all outcomes and to maintain compactness of our type spaces which is necessary for our results. In Section 8.1, we discuss how the worst outcome assumption can be relaxed while maintaining non-triviality and compactness of preferences.

Our results are closely tied to a number of existing literatures. Most importantly, Abreu and Matsushima (1992b) characterize (full) virtual Bayesian implementability of social choice functions for a finite type space under the solution concept of iterated deletion of strictly dominated strategies. A necessary condition is a "measurability" condition that, in the language of this paper, requires that the social choice function gives the same outcome to strategically indistinguishable types. They provide a characterization of the measurability condition that essentially states that types are strategically distinguishable if and only if they differ in their preference hierarchies. Iterated deletion of strictly dominated strategies is equivalent, in their setting, to a refined version of rationalizability—interim correlated rationalizability—that is intermediate between equilibrium and IPCR. They also show that the measurability condition is necessary for virtual Bayesian implementation in equilibrium, and so their argument establishes a characterization of strategic dis-

[1] This is discussed further in Section 8.6.

tinguishability for equilibrium as well. Given that our preference hierarchy is an infinite space, our revealing mechanism provides a generalization of the result of Abreu and Matsushima (1992b) to infinite type spaces. As well as raising new technical challenges, a benefit of the extension is that the equivalence relation between preference hierarchies and strategic distinguishability can be stated in terms of a universal space and thus without reference to a specific type space from which the types are drawn.[2]

As we noted above, our universal interdependent preference space construction is mathematically equivalent to the construction of the universal belief space of Mertens and Zamir (1985), although we are giving it a quite different interpretation. Epstein and Wang (1996) construct a universal space of hierarchies of non-expected utility preferences, incorporating non-expected utility preferences such as ambiguity aversion, but maintaining monotonicity as well as additional regularity conditions. We must dispense with monotonicity to incorporate the interdependence of preferences we want to capture. We relax monotonicity to the worst outcome assumption, but impose independence to get an expected utility representation. Di Tillio (2008) allows general preferences, and thus does not require Epstein and Wang's monotonicity condition or independence, but restricts attention to preferences over finite outcomes at every level of the hierarchy.[3]

A number of authors have considered problems that arise in behaviorally identifying psychologically motivated properties of preferences that involve interdependence (see Levine (1998) and Weibull (2004)) such as conditional altruism (e.g. I want to be generous only to those people who are generous themselves). Motivated by such problems, Gul and Pesendorfer (2007) construct a universal space of interdependent preference types. We construct a different universal interdependent preference space. They identify a maximal set of types which captures all distinctions that can be expressed in a natural language. When they consider applications of their universal space to incomplete information settings, they treat incomplete information separately and thus they do not address the interaction (and indistinguishability in a state dependent expected utility setting) of beliefs and utilities. Our focus is on static games and solution concepts (equilibrium and rationalizability) without sequential rationality or other refinements of those solution concepts. This implies that, in a complete information setting, it is not possible to identify any interdependence in agents' types (a point emphasized in our leading example of Section 3). Thus our universal space of interdependent types ends up being much coarser than that of Gul and Pesendorfer (2007). In

[2] See Section 8.4 for a brief discussion of how our results might be used to extend the implementation results of Abreu and Matsushima (1992b) to infinite type spaces. In Section 8.5, we discuss how the analysis in this paper is related to Bergemann and Morris (2009), which showed that robust virtual implementation is possible only if there is not too much interdependence in preferences.

[3] See Section 8.2 for a brief discussion of how our results might change if we dropped the expected utility assumption.

particular, their types reflect much counterfactual information (what preferences would be conditional on other agents' types) that cannot be strategically distinguished in our setting, with static games and solution concepts.[4]

A recent literature (Dekel, Fudenberg and Morris [DFM] (2006, 2007), Ely and Pęski (2006), Liu (2009), Sadzik (2010)) has examined what can be learned about agents' beliefs and higher-order beliefs about a state space Θ when it is (informally) assumed that there is common certainty of agents' "payoffs" as a function of their actions in a game and the realized state $\theta \in \Theta$. Our results can be understood as a relaxation of the assumption of common certainty of payoffs in that literature. In particular, that literature can be summarized as follows. DFM show that two types have the same interim correlated rationalizable (ICR) actions if and only if they have the same higher-order beliefs, i.e., they map to the same Mertens and Zamir [MZ] (1985) type. Thus, in the language of this paper, MZ types characterize strategic equivalence for ICR under the common certainty of payoffs assumption. ICR is a permissive solution concept that allows agents' actions to reveal information about others' actions and the payoff relevant state. If restrictions are put on what can be revealed, as in the notion of interim independent rationalizability (IIR) of DFM (2007), then finer distinctions over types are required to characterize strategic equivalence. Ely and Pęski (2006) describe richer hierarchies than MZ types which characterize IIR in two agent games. Liu (2009) and Sadzik (2009) discuss even richer information needed to characterize Bayesian Nash equilibrium (BNE). Although not highlighted in this literature, it is easy to deduce from these existing results that MZ types characterize strategic indistinguishability for all three solution concepts (ICR, IIR and BNE); in other words, two types have an ICR/IIR/BNE action in common in every mechanism if and only if they have the same MZ type. To see why, note that we can always find a BNE action they have in common by looking for pooling equilibria where redundant information is ignored. Thus a summary of the "common certainty of payoffs" literature is:

	strategically equivalent	strategically indistinguishable
ICR	Mertens-Zamir space	Mertens-Zamir space
IIR	Ely-Pęski space	Mertens-Zamir space
BNE	richer Liu/Sadzik space	Mertens-Zamir space

Our results in this paper offer a clean generalization of this picture. This literature combines beliefs and higher-order beliefs about some payoff relevant states with common certainty of a mapping from action profiles and payoff relevant states to payoffs. Relaxing the common certainty of payoffs assumption, we must construct a universal space of higher-order (expected utility) preferences. We

[4] See Section 8.3 for a brief discussion of how results might change with dynamic games and solution concepts incorporating sequential rationality.

show that this characterizes strategic indistinguishability for equilibrium, for IPCR and for any solution concept in between. We show that it also characterizes strategic equivalence for IPCR but not necessarily for more refined versions of rationalizability and equilibrium.

The paper is organized as follows. Section 2 describes our setting and poses the strategic distinguishability question for equilibrium. Section 3 considers in detail an interdependent preferences example to motivate the approach and results in the paper. Section 4 describes the construction of the universal space of interdependent preferences. Section 5 reports our main result: our universal space characterizes equilibrium strategic distinguishability. Section 6 introduces the solution concept of interim preference correlated rationalizability, and presents the proof that our universal space characterizes strategic distinguishability for equilibrium, IPCR and everything in between. Section 7 formally introduces the finer strategic equivalence relation, shows that our universal space characterizes IPCR strategic equivalence and discusses the formal connection with the common certainty of payoffs literature. Section 8 concludes.

2 The Setting and Benchmark Question

An outside observer will see a finite set of agents, $\mathcal{I} = \{1, ..., I\}$, making choices in strategic situations, where there is a finite set of outcomes Z and a compact and metrizable set of observable states Θ. We will maintain the assumption that, for each agent i, there is an outcome $w_i \in Z$ which is a worst outcome for that agent; in Section 8.1, we discuss relaxations of this assumption.

We are interested in what the outside observer can infer about agents' (perhaps interdependent) preferences by observing agents' rational choices in strategic situations. We will consider standard Harsanyi type space models of agents' perhaps interdependent preferences. A type space consists of a measurable set of unobservable states, Ω, and for each agent i, a measurable space of types T_i, a measurable belief function $\nu_i \colon T_i \to \Delta(\Theta \times \Omega \times T_{-i})$ and a bounded and measurable utility function $u_i \colon \Theta \times \Omega \times T \times Z \to \mathbb{R}$. Consistent with the assumption that agent i has a worst outcome $w_i \in Z$, we require

$$u_i(\theta, \omega, t, z) \geq u_i(\theta, \omega, t, w_i)$$

for all $\theta \in \Theta$, $\omega \in \Omega$, $t \in T$ and $z \in Z$. In addition, we will make the non-triviality assumption that for every $t_i \in T_i$ and $\nu_i(\cdot\,|t_i)$-almost every $(\theta, \omega, t_{-i}) \in \Theta \times \Omega \times T_{-i}$, there exists some $z \in Z$ such that $u_i(\theta, \omega, t, z) > u_i(\theta, \omega, t, w_i)$. Thus a Harsanyi type space is given by $\mathcal{T} = \left(\Omega, (T_i, \nu_i, u_i)_{i \in \mathcal{I}}\right)$.

We define a belief-closed subset of the type space to be a product set of agents' types where each agent is sure to be in that subset. Formally, a product set $\tilde{T} = \prod_i \tilde{T}_i$ of types with measurable $\tilde{T}_i \subseteq T_i$ is belief-closed if for every $i \in \mathcal{I}$ and $t_i \in \tilde{T}_i$, $\nu_i\left(\Theta \times \Omega \times \tilde{T}_{-i}\,|t_i\right) = 1$.

A strategic situation is modelled as a mechanism, where each agent i has a finite set of actions A_i and an outcome function $g\colon \Theta \times A \to \Delta(Z)$. Thus a mechanism is defined by $\mathcal{M} = \left((A_i)_{i\in\mathcal{I}}, g\right)$.

The pair $(\mathcal{T}, \mathcal{M})$ describes a game of incomplete information. A strategy for agent i in this game is a measurable function $\sigma_i\colon T_i \to \Delta(A_i)$. We extend the domain of g to mixed strategies in the usual way. Bayesian Nash equilibria do not always exist on large type spaces. However, even when equilibria do not exist on large type spaces, equilibria may exist on belief-closed subsets of the large type space. We will follow Sadzik (2010) in defining such "local" equilibria.

Definition 1 *A strategy profile $\sigma = (\sigma_i)_{i\in\mathcal{I}}$ is a local equilibrium of the game $(\mathcal{T}, \mathcal{M})$ on the belief-closed subspace $\tilde{T}$ if, for every $i \in \mathcal{I}$ and $t_i \in \tilde{T}_i$, $\sigma_i(t_i)$ maximizes*

$$\int_{\Theta\times\Omega\times T_{-i}} u_i(\theta, \omega, (t_i, t_{-i}), g(\theta, (a_i, \sigma_{-i}(t_{-i})))) d\nu_i(t_i)(\theta, \omega, t_{-i}).$$

Let $E_i(t_i, \mathcal{T}, \mathcal{M})$ be the set of all local equilibrium actions of type t_i, i.e., the set of actions played with positive probability by t_i in any local equilibrium of $(\mathcal{T}, \mathcal{M})$ on any belief-closed subspace $\tilde{T}$ with $t_i \in \tilde{T}_i$.

We say that a type t_i is countable if there exists a countable belief-closed subspace $\tilde{T} = \prod_j \tilde{T}_j$ with $t_i \in \tilde{T}_i$. By Kakutani's fixed-point theorem, $E_i(t_i, \mathcal{T}, \mathcal{M}) \neq \emptyset$ if t_i is countable.

The main relation between types that we seek to characterize in this paper is the following.

Definition 2 *Two types of agent i, $t_i \in \mathcal{T}$ and $t_i' \in \mathcal{T}'$, are strategically indistinguishable if, for every mechanism $\mathcal{M}$, there exists some action that can be chosen by both types, so that*

$$E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i(t_i', \mathcal{T}', \mathcal{M}) \neq \emptyset$$

for every $\mathcal{M}$. Conversely, t_i and t_i' are strategically distinguishable if there exists a mechanism in which no action can be chosen by both types, so that

$$E_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap E_i(t_i', \mathcal{T}', \mathcal{M}^*) = \emptyset$$

for some $\mathcal{M}^*$.

Our main result will be a characterization of strategic distinguishability. Before reporting our result, we report examples to motivate and provide intuition for results.

3 Examples and Motivation

In this section, we illustrate by means of an example, that there are many equivalent ways using a Harsanyi type space to describe interdependent preference, which will all give rise to indistinguishable behavior. We refer to this multiplicity in the representation of interdependent preferences as redundancy. Our purpose in analyzing the example is to describe the redundancy, use it to motivate a canonical - hierarchical - representation of interdependent types, and give an intuition why this representation exactly captures strategic distinguishability as described in the previous section. We begin with the elementary issue of *decision-theoretic redundancy* - those redundancies that would already arise in a single person decision problem - and then discuss *strategic redundancy* - more subtle redundancies that arise from the interdependence of preferences.

3.1 Decision Theoretic Redundancy

Two detectives, 1 and 2, must decide on the guilt or innocence of a suspect. There are three possible states: the suspect is innocent (probability $\frac{1}{3}$), the suspect committed the crime in the morning (probability $\frac{1}{3}$), and the suspect committed the crime in the afternoon (probability $\frac{1}{3}$). Each detective observes an alibi; if the suspect is innocent, each alibi is equally likely for each detective; if the suspect is guilty, the alibi is for a time different from when the crime was committed. Detective 1 always remembers his alibi correctly, but, if the suspect is guilty, detective 2 remembers a morning alibi correctly but mis-remembers an afternoon alibi with probability ε - assumed strictly greater than 0 for now. There are three possible outcomes: conviction, acquittal or no verdict. We assume no verdict or a "wrong" decision give utility 0 to the detectives, while a correct decision gives utility 1.

This scenario can be described with a Harsanyi type space as follows. There are no observed states. The unobserved states, $\Omega = \{\omega_I, \omega_M, \omega_A\}$, correspond to innocent, morning crime and afternoon crime respectively. The type spaces are $T_1 = T_2 = \{m, a\}$, corresponding to morning and afternoon alibis respectively. Beliefs are generated by a common prior over states and types represented in the following tables:

$\omega = \omega_I$:

$t_1 \backslash t_2$	m	a
m	$\frac{1}{12}$	$\frac{1}{12}$
a	$\frac{1}{12}$	$\frac{1}{12}$

$\omega = \omega_M$:

$t_1 \backslash t_2$	m	a
m	0	0
a	$\frac{\varepsilon}{3}$	$\frac{1-\varepsilon}{3}$

$\omega = \omega_A$:

$t_1 \backslash t_2$	m	a
m	$\frac{1}{3}$	0
a	0	0

Outcomes are $Z = \{C, A, N\}$, corresponding to convict, acquit and no verdict respectively, and

utility functions are

$$u_1(\omega,t,z)=u_2(\omega,t,z)=\begin{cases}1, & \text{if } (\omega,z)=(\omega_M,C),\ (\omega_A,C) \text{ or } (\omega_I,A);\\ 0, & \text{if otherwise.}\end{cases}$$

Note that example has the special features that there are no private goods and identical interests. This simplifies the example but plays no role in the general analysis.

While this is one formal representation, there are many equivalent ways of describing a type's beliefs and utilities that give rise to the same individual preferences and thus behavior. We refer to this as decision theoretic redundancy. A first simple and well known observation is that states that are not observed by any agent are redundant and can be integrated out (see, for example, Milgrom (2004), page 159, for a discussion). Thus an alternative Harsanyi type space representation of the above example is the following. There are no unobservable states, the type spaces remain $T_1=T_2=\{m,a\}$; but the utility functions have the form:

$t_1\backslash t_2$	m	a
m	$\frac{4}{5},\frac{1}{5}$	$0,1$
a	$\frac{4\varepsilon}{1+4\varepsilon},\frac{1}{1+4\varepsilon}$	$\frac{4-4\varepsilon}{5-4\varepsilon},\frac{1}{5-4\varepsilon}$

(1)

where the entries refer to the (expected) utility from conviction and acquittal (the utility of no verdict is always 0); and beliefs on $T_1\times T_2$ consistent with the following common prior:

	m	a
m	$\frac{5}{12}$	$\frac{1}{12}$
a	$\frac{1+4\varepsilon}{12}$	$\frac{5-4\varepsilon}{12}$

(2)

While the interdependence in the original example has an informational motivation, note that once we integrate out the unobserved states, we can just as well provide a psychological interpretation. For example, suppose that the detectives are unconcerned about the guilt or innocence of the suspect, and the signals refer to whether each detective belongs to tribe m or tribe a. The detectives get a kick out of convicting the defendant when they are from the same tribe, but prefer acquittal when they come from different tribes. This behavioral story is also represented by this type space. This example illustrates the elementary but important observation that - in our setting - there is no way of telling informational and psychological explanations of preference interdependence apart.

Another form of decision theoretic redundancy in the description of Harsanyi types is that since the utility function is "state-dependent," i.e., is allowed to depend on types, the distinction between "utility" and "beliefs" is arbitrary and all we can observe is the product of the two. Another way of making this point is to observe that the choice of numeraire is arbitrary but affects whether

interdependence is reflected in beliefs or utilities. We can illustrate this with another equivalent representation of the above example. Let beliefs be generated by a uniform and independent common prior over type profiles

$t_1 \backslash t_2$	m	a
m	$\frac{1}{4}$	$\frac{1}{4}$
a	$\frac{1}{4}$	$\frac{1}{4}$

(3)

and let the utility function be

$t_1 \backslash t_2$	m	a
m	$4, 1$	$0, 1$
a	$4\varepsilon, 1$	$4 - 4\varepsilon, 1$

(4)

where again the entries refer to the (expected) utility from conviction and acquittal respectively. Observe that, for each type profile, the product of the prior belief (in table 3) and utility (in table 4) in this representation is equal to (3 times) the product of the prior belief (in table 2) and the utility (in table 1) in the previous representation, and thus these constitute representations of the same individual preferences. These re-normalizations are possible because the state dependent utility representations of the expected utility preferences do not pin down the probabilities. This well-known fact is discussed, for example, in Myerson (1991) where he labels two incomplete information games where one is such a re-normalization of the other as representing "fully equivalent games" - and it is well known and relevant for empirical auction research (see Paarsch and Hong (2006)).

Our solution to these two forms of decision theoretic redundancy (integrating out unobserved states and inseparability of beliefs and utilities) will be to work with preference type spaces, where unobserved states are integrated out and types are identified with preferences over Anscombe-Aumann acts contingent on observable states and others' types. Thus we will abstract from numeraires, beliefs and utilities in the preference type space representation. In the example, the preferences of type m of detective 1 are summarized by the observation that he will maximize 4 times the probability of conviction if detective 2 is type m plus the probability of acquittal if detective 2 is type m plus the probability probability of acquittal if detective 2 is type a. Formally, given a choice between the acts $f : T_2 \to \Delta(Z)$ and $f' : T_2 \to \Delta(Z)$, he will weakly prefer f to f' if

$$4f(m)(C) + f(m)(A) + f(a)(A) \geq 4f'(m)(C) + f'(m)(A) + f'(a)(A).$$

Mapping Harsanyi type spaces into preference type spaces is straightforward. We give a general description of the transformation in Section 4.2. However, while the preference type spaces remove decision theoretical redundancies, they do not provide a "natural" language to discuss interdependent preferences, since they are self-referential. Nor do they provide a characterization of strategic distinguishability. We will therefore introduce a natural canonical way to represent interdependent

types in Section 4.3. We can illustrate this construction, and its relevance for strategic distinguishability, with the example. Consider first the detectives' "unconditional" or "first level" preference over "unconditional lotteries," i.e., constant acts that do not depend on the other detective's type. Both types of both detectives strictly prefer any verdict to no verdict. Thus "no verdict" is a worst outcome for both detectives. The first level preference of any type is then characterized by his "marginal rate of substitution" between conviction and acquittal, i.e., the rate at which he is willing to exchange probability of acquittal for probability of conviction. Thus type m of detective 2 has a "marginal rate of substitution" $2(1+\varepsilon)$ between conviction and acquittal, i.e., he would be indifferent increasing the probability of conviction by δ or increasing the probability of acquittal by $2(1+\varepsilon)\delta$. On the other hand, type a of detective 2 will have an unconditional preference with marginal rate of substitution $2(1-\varepsilon)$ between conviction and acquittal. Thus we could distinguish detective 2's type from his first level preferences alone, and thus identify his preferences from his behavior in a single person decision problem alone. But both types of detective 1 have identical first level preferences, with an marginal rate of substitution of 2 independent of the signal observed. Thus detective 1's types cannot be distinguished by their first level types and thus could not be distinguished in a single agent decision problem. However, if detective 1 is type m, then conditional on detective 2's unconditional odds ratio being $2(1+\varepsilon)$, his conditional marginal rate of substitution between conviction and acquittal - is 4, but conditional on detective 2's unconditional preference being $2(1-\varepsilon)$, his conditional marginal rate of substitution between conviction and acquittal is 0. These conditional preferences are part of type m of detective 1's second level preferences. In this example, second level preferences contain enough information to strategically distinguish types.

In Section 4, we provide a formal description of a universal space of possible expected utility types, consisting of (i) unconditional (expected utility) preferences; (ii) preferences conditional on others' unconditional preferences; and so on. In Section 5, we confirm that two types are guaranteed to behave differently in equilibrium of some mechanism if and only if they correspond to different types in this universal space (Theorem 1).

But before we move to the general analysis, we will give another example demonstrating how two types that may look quite different in a preference type space, and are decision theoretically distinct, map to the same preference hierarchy in the universal type space. We refer to this phenomenon as strategic redundancy. We will then use this example to motivate our later results concerning strategic equivalence and alternative definitions of rationalizability.

3.2 Strategic Redundancy

In our example, we assumed that detective 2 mis-remembered an afternoon alibi with probability $\varepsilon > 0$. This assumption made the example asymmetric between detectives 1 and 2 and ensured that detective 2's types could be distinguished by their first level preferences. Now consider what happens if we restore symmetry by setting $\varepsilon = 0$. Consider the uniform prior representation of the Harsanyi space (tables 3 and 4); if $\varepsilon = 0$, the payoffs table becomes:

$t_1 \backslash t_2$	m	a
m	$4, 1$	$0, 1$
a	$0, 1$	$4, 1$

Now both types of both detectives have marginal rate of substitution 2 between conviction and acquittal. Thus both types of both detectives have common certainty that each detectives has marginal rate of substitution 2. So both types are equivalent (in terms of their preference hierarchy) to complete information types with utilities $(2, 1, 0)$ for conviction, acquittal and no verdict respectively. We say that type m and a are "redundant types" - following the terminology of Mertens and Zamir (1985). This *strategic redundancy* is different from but related to the redundancy in Mertens and Zamir (1985), Ely and Pęski (2006) and Dekel, Fudenberg and Morris (2007). We discuss the connection in detail in Section 7.3

The example highlights the simple but important point that in our universal preference hierarchy space, types can exhibit interdependent preference only if there is not complete information. With complete information, there is common certainty of each agent's preferences and any interdependence in the agents' minds will not necessarily be reflected in their behavior and so cannot be strategically distinguished.

However, it is easy to construct a mechanism where equilibrium actions of one type are not equilibrium actions of the other type. Consider the mechanism where each detective either makes a report or "opts out". If either detective opts out, the suspect is convicted. If neither detective opts out and both announce the same type, the suspect is convicted with probability δ and there is no verdict with probability $1 - \delta$; if neither detective opts out and they announce different types, the suspect is acquitted with probability δ and there is no verdict with probability $1 - \delta$. The mechanism is summed up in the following table where the triple in each box corresponds to the lottery over outcomes convict, acquit and no verdict, respectively:

$t_1 \backslash t_2$	`m`	`a`	`optout`
`m`	$(\delta, 0, 1-\delta)$	$(0, \delta, 1-\delta)$	$(1, 0, 0)$
`a`	$(0, \delta, 1-\delta)$	$(\delta, 0, 1-\delta)$	$(1, 0, 0)$
`optout`	$(1, 0, 0)$	$(1, 0, 0)$	$(1, 0, 0)$

(5)

If $\delta > \frac{4}{5}$, and this mechanism is played with the original Harsanyi type space with redundant types, there is a strict equilibrium where the detectives "tell the truth," i.e., type m sends message `m` and type a sends message `a`. Each type's expected utility is $\frac{5}{2}\delta > 2$, while the expected utility from `optout` is 2. But if $\delta < 1$, the unique equilibrium for complete information types has each detective opting out (giving expected utility 2).

In Section 7, we will introduce a formal definition of strategic equivalence: two types are strategically equivalent if they have the same set of equilibrium actions in any mechanism. We have just shown that types m and a in our Harsanyi type space and the complete information types are strategically indistinguishable but not strategically equivalent. The gap arises because of the existence of multiple equilibria: in any mechanism, there is always an equilibrium on the Harsanyi type space where types m and a pool, i.e., choose the same action, and the resulting equilibrium corresponds to an equilibrium in the complete information game. This establishes their strategic indistinguishability. But in the case of the mechanism we described above, there is another equilibrium where the redundant types behave in a way that the complete information types never could.

This example illustrates that preference hierarchies do not contain enough information to characterize the strategic equivalence of Harsanyi types. However, it turns out that the preference hierarchy does characterize strategic equivalence for (one version of) rationalizability (Theorem 3). But this result is sensitive to the exact definition of rationalizability. We can illustrate this also in the example. It seems natural to argue that `optout` strictly dominates action `m` or action `a` in the example, and therefore should be the only rationalizable action for the complete information types; and `optout` is the only *interim correlated rationalizable* (ICR) action, in the sense of Dekel, Fudenberg and Morris (2007), for the complete information types. Since any equilibrium action must be interim correlated rationalizable, the types are again strategically indistinguishable but not strategically equivalent, if we use ICR instead of equilibrium in the definition of those concepts.

But there are subtleties in defining rationalizable outcomes. In the solution concept of ICR, each detective is allowed to have conjectures in which his opponent's actions and observable states are correlated in his mind, so that the opponent's action reveals information about the observable state in the detective's mind. Analogously, in our context, it is natural to allow detectives to believe that the other's action will reveal information about their own preferences. In Section 6, we will formally describe a generalization of ICR, called *interim preference correlated rationalizability* (IPCR), where we require detectives' preferences unconditional on the opponent's action to respect the detective's preference hierarchy, but allow any preferences contingent on the opponent's action with the correct marginal on unconditional preferences. We show that two types are strategically equivalent under this solution concept if and only if they map to the same type in the universal space

of interdependent preferences. However, for any more refined solution concepts (such as equilibrium and interim correlated rationalizability), strategical equivalence generates a finer partition than our universal space.

We can illustrate this with our example. Consider a complete information type, with common certainty that both detectives' marginal rates of substitution is 2. Suppose that each detective believed that there was correlation between the other detective's action and the guilt of the suspect. If he believed there was positive correlation, it would make sense for the complete information type to choose the same action (in the mechanism described in table 5) as the action he thinks the other detective is most likely to choose, while if there was negative correlation he would pick the other action. Thus both actions are IPCR. The example illustrates that the solution concept is very permissive. Intuitively, the solution concept of IPCR allows rational detectives to build into their preferences any "redundant" elements in the Harsanyi type space, and thus the redundant elements do not matter.

4 Preference Types

We introduce preference type spaces that capture interdependent preferences and have no decision theoretic redundancy. We then construct a universal preference type space, which consists of preference hierarchies.

4.1 State-Dependent Preferences

We first define state-dependent preferences for a single agent in the framework of Anscombe and Aumann (1963). We begin with a measurable space X of states and a finite set Z of outcomes with $|Z| \geq 2$. An (*Anscombe-Aumann*) *act* is a measurable mapping from X to $\Delta(Z)$. The set of all such acts is denoted by $F(X)$ and endowed with the sup norm. For $y, y' \in \Delta(Z)$ and measurable $E \subseteq X$, $y_E y'$ is the act that yields the lottery y over E and the lottery y' over $X \setminus E$. We consider the following conditions on binary relation $\succsim$ over $F(X)$. For a fixed worst outcome $w \in Z$, we define $P_w(X)$ to be the set of all binary relations over $F(X)$ that have a non-trivial state-dependent expected utility representation respecting the worst outcome:

Definition 3 *A binary relation over $F(X)$ is a (worst outcome w) expected utility preference if there exists $\mu \in \Delta(X \times (Z \setminus \{w\}))$ that satisfies*

$$f \succsim f' \Leftrightarrow \int_{X\times(Z\setminus\{w\})} f(x)(z)d\mu(x,z) \geq \int_{X\times(Z\setminus\{w\})} f'(x)(z)d\mu(x,z)$$

for any $f, f' \in F(X)$.

This representation can be axiomatized with a simple variant of standard arguments in decision theory.

1. *completeness*: for every $f, f' \in F(X)$, $f \succsim f'$ or $f' \succsim f$.
2. *transitivity*: for every $f, f', f'' \in F(X)$, if $f \succsim f'$ and $f' \succsim f''$, then $f \succsim f''$.
3. *independence*: for every $f, f', f'' \in F(X)$ and $\lambda \in (0,1]$, $f \succsim f'$ if and only if $\lambda f + (1-\lambda) f'' \succsim \lambda f' + (1-\lambda) f''$.
4. *continuity*: for every $f, f', f'' \in F(X)$, if $f \succ f' \succ f''$, then there exists $\varepsilon \in (0,1)$ such that $(1-\varepsilon) f + \varepsilon f'' \succ f' \succ (1-\varepsilon) f'' + \varepsilon f$.
5. *monotone continuity*: for every $z, z', z'' \in Z$ with $z \succ z'$ and decreasing sequence $\{E_n\}_{n \in \mathbb{N}}$ of measurable subsets of X with $\bigcap_n E_n = \emptyset$, there exists $n \in \mathbb{N}$ such that ${z''}_{E_n} z \succ z'$ and $z \succ {z''}_{E_n} z'$.
6. *non-triviality*: there exist $f, f' \in F(X)$ with $f \succ f'$.
7. *worst outcome* w: for every $f \in F(X)$, $f \succsim w$.

Proposition 1 *$\succsim \in P_w(X)$ if and only if it satisfies completeness, transitivity, independence, continuity, monotone continuity, non-triviality and w worst outcome.*

An event $E \subseteq X$ is *$\succsim$-null* if $z_E w \sim w$ for every $z \in Z$. For $\succsim$ represented by $\mu \in \Delta(X \times (Z \setminus \{w\}))$, E is $\succsim$-null if and only if $\mu(E \times (Z \setminus \{w\})) = 0$. An event E is *$\succsim$-certain* if $X \setminus E$ is $\succsim$-null.

For a preference $\succsim \in P_w(X)$ and a measurable space Y, a measurable mapping $\varphi \colon X \to Y$ induces a preference $\varphi^P(\succsim) \in P_w(Y)$ given by

$$f \; \varphi^P(\succsim) \; f' \Leftrightarrow f \circ \varphi \succsim f' \circ \varphi$$

for any $f, f' \in F(Y)$. In particular, for a preference $\succsim \in P_w(X \times Y)$, the projection from $X \times Y$ to X induces the *marginal preference of* $\succsim$, $\mathrm{mrg}_X \succsim \in P_w(X)$, which is the restriction of $\succsim$ to acts over $X \times Y$ that do not depend on the Y-coordinate.

$P_w(X)$ is treated as a measurable space with the σ-algebra generated by $\{\succsim \in P_w(X) \mid f \succsim f'\}$ for any $f, f' \in F(X)$. If X is a topological space, then $P_w(X)$ is also endowed with the weak topology generated by $\{\succsim \in P_w(X) \mid f \succ f'\}$ for any continuous $f, f' \in F(X)$.

We will sometimes work with redundant representations of state-dependent preferences in which we distinguish between beliefs and utilities. For a belief $\nu \in \Delta(X)$ and a bounded and measurable utility function $u \colon X \times Z \to \mathbb{R}$ with $u(x,z) \geq u(x,w)$ for all $x \in X$ and $z \in Z$, with strict

inequalities for ν-almost every $x \in X$ and some $z \in Z$, we write $\succsim^{\nu,u} \in P_w(X)$ for the induced preference, i.e.,

$$f \succsim^{\nu,u} f' \Leftrightarrow \int_X u\left(x, f(x)\right) d\nu(x) \geq \int_X u\left(x, f'(x)\right) d\nu(x)$$

for any $f, f' \in F(X)$.

4.2 Preference Type Spaces

Fix a finite set $\mathcal{I} = \{1, \ldots, I\}$ of agents with $I \geq 2$ and a compact and metrizable set Θ of states of nature. Each agent i has the worst outcome $w_i \in Z$. We write $P_i(X) \equiv P_{w_i}(X)$.

Definition 4 *A preference type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ consists of, for each $i \in \mathcal{I}$, a measurable space T_i of agent i's types and a measurable mapping $\pi_i \colon T_i \to P_i(\Theta \times T_{-i})$ that maps his types to preferences over acts over observable states and his opponents' types, where $T_{-i} = \prod_{j \in \mathcal{I} \setminus \{i\}} T_j$.*

Similarly to Harsanyi type spaces, a product $\tilde{T} = \prod_i \tilde{T}_i$ of measurable sets $\tilde{T}_i \subseteq T_i$ is preference-closed if for every $i \in \mathcal{I}$ and $t_i \in \tilde{T}_i$, $\Theta \times \tilde{T}_{-i}$ is $\pi_i(t_i)$-certain. A type t_i is countable if there exists a countable preference-closed subspace $\tilde{T} = \prod_j \tilde{T}_j$ with $t_i \in \tilde{T}_i$.

For a given Harsanyi type space $\mathcal{T} = \left(\Omega, (T_i, \nu_i, u_i)_{i \in \mathcal{I}}\right)$, we have observed in Section 3.1 two forms of decision theoretic redundancy: first, we can integrate out unobserved states; second, the distinction between beliefs and utilities is not relevant. In particular, a type t_i of agent i is characterized in the Harsanyi type space by a belief $\nu_i\left(t_i\right) \in \Delta(\Theta \times \Omega \times T_{-i})$ and a utility function $u_i\left(t_i\right) : \Theta \times \Omega \times T_{-i} \times Z \to \mathbb{R}$. Together, they induce the preference relation

$$\pi_i^{\nu_i, u_i}\left(t_i\right) \equiv \mathrm{mrg}_{\Theta \times T_{-i}} \succsim^{\nu_i(t_i), u_i(t_i)}$$

over $F\left(\Theta \times T_{-i}\right)$. Thus the preference type space $\mathcal{T} = (T_i, \pi_i^{\nu_i, u_i})_{i \in \mathcal{I}}$ embodies decision theoretically non-redundant information in the Harsanyi type space, and we will abuse notation by writing $\mathcal{T}$ for both when no confusion arises. We will refer to $(T_i, \pi_i^{\nu_i, u_i})_{i \in \mathcal{I}}$ as the preference type space induced by Harsanyi type space $\left(\Omega, (T_i, \nu_i, u_i)_{i \in \mathcal{I}}\right)$ and refer to types t_i as belonging to both a Harsanyi type space and its induced preference-type space.

4.3 The Universal Preference Type Space

We now construct the universal preference type space à la Mertens and Zamir (1985) and Brandenburger and Dekel (1993). In light of the isomorphism between preferences $P_i(X)$ and probability measures $\Delta(X \times (Z \setminus \{w_i\}))$ that represent them, this is straightforward and we report standard results with minimal comments.

Let $X_{i,0} = \{*\}$ be initialized with a single element, and let $X_{i,n} = X_{i,n-1} \times P_i(\Theta \times X_{-i,n-1})$ for each $n \geq 1$. Note that $X_{i,n} = \prod_{k=0}^{n-1} P_i(\Theta \times X_{-i,k})$. Let $X_{i,\infty} = \prod_{n=0}^{\infty} P_i(\Theta \times X_{-i,n})$. Each $X_{i,n}$ is compact and metrizable, and thus $X_{i,\infty}$ is compact and metrizable. Let $Y_{i,0} = \prod_{n=0}^{\infty} \Delta(\Theta \times X_{-i,n} \times (Z \setminus \{w_i\}))$ be the set of hierarchies of probability measures for agent i. A hierarchy of probability measures, $\{\mu_{i,n}\}_{n=1}^{\infty} \in Y_{i,0}$, is *coherent* if $\mathrm{mrg}_{\Theta \times X_{-i,n-2} \times (Z \setminus \{w_i\})} \mu_{i,n} = \mu_{i,n-1}$ for every $n \geq 2$. Let $Y_{i,1} \subset Y_{i,0}$ be the set of all coherent hierarchies of probability measures.

For each $\mu_{i,n} \in \Delta(\Theta \times X_{-i,n-1} \times (Z \setminus \{w_i\}))$ with $n \geq 1$, let $\rho_{i,n}(\mu_{i,n}) \in P_i(\Theta \times X_{-i,n-1})$ denote the preference represented by $\mu_{i,n}$. Let $\rho_i \colon Y_{i,0} \to X_{i,\infty}$ be the collection of such mappings $\rho_{i,n}$. Similarly, for each $\mu_{i,\infty} \in \Delta(\Theta \times X_{-i,\infty} \times (Z \setminus \{w_i\}))$, let $\rho_{i,\infty}(\mu_{i,\infty}) \in P_i(\Theta \times X_{-i,\infty})$ denote the preference represented by $\mu_{i,\infty}$.

By the Kolmogorov extension theorem, there is a homeomorphism $\psi_i \colon Y_{i,1} \to \Delta(\Theta \times X_{-i,\infty} \times (Z \setminus \{w_i\}))$. Let $T_{i,1} = \rho_i(Y_{i,1}) \subset X_{i,\infty}$. Note that every $\{\succsim_{i,n}\}_{n=1}^{\infty} \in T_{i,1}$ satisfies coherency, i.e., $\mathrm{mrg}_{\Theta \times X_{-i,n-2}} \succsim_{i,n} = \succsim_{i,n-1}$ for every $n \geq 2$. We convert ψ_i to a mapping between preference spaces and obtain a homeomorphism $\psi_{i,P} = \rho_{i,\infty} \circ \psi_i \circ \rho_i^{-1} \colon T_{i,1} \to P_i(\Theta \times X_{-i,\infty})$.

For $n \geq 2$, let

$$T_{i,n} = \{t_i \in T_{i,1} \mid \Theta \times T_{-i,n-1} \text{ is } \psi_{i,P}(t_i)\text{-certain}\}$$

and $T_i^* = \bigcap_{n=1}^{\infty} T_{i,n}$. Note that $T_{i,n}$ is compact for every $n \geq 1$, and hence T_i^* is also compact. Thus we obtain a homeomorphism $\pi_i^* = \psi_{i,P}|_{T_i^*} \colon T_i^* \to P_i(\Theta \times T_{-i}^*)$. We call $\mathcal{T}^* = (T_i^*, \pi_i^*)_{i \in \mathcal{I}}$ the *universal preference type space*.

Definition 5 *For two preference type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ and $\mathcal{T} = (T_i', \pi_i')_{i \in \mathcal{I}}$, a profile $(\varphi_i)_{i \in \mathcal{I}}$ of measurable mappings $\varphi_i \colon T_i \to T_i'$ preserves preferences if*

$$\pi_i' \circ \varphi_i = (\mathrm{id}_\Theta \times \varphi_{-i})^P \circ \pi_i$$

for every $i \in \mathcal{I}$.

Fix a preference type space $\mathcal{T} = (T_i, \pi_i)_{i=1,2}$. For each type $t_i \in T_i$ of agent i, let $\hat{\pi}_{i,1}(t_i) = \mathrm{mrg}_\Theta \pi_i(t_i)$ and $\hat{\pi}_{i,n}(t_i) = (\mathrm{id}_\Theta \times (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1}))^P(\pi_i(t_i))$ for each $n \geq 2$. Each $\hat{\pi}_{i,n}(t_i)$ denotes the *n-th order preference of t_i*, and $\hat{\pi}_i(t_i) = \{\hat{\pi}_{i,n}(t_i)\}_{n=1}^{\infty}$ the *hierarchy of preferences of t_i*. For any Harsanyi type space, $\mathcal{T} = \left(\Omega, (T_i, \nu_i, u_i)_{i \in \mathcal{I}}\right)$ and $t_i \in T_i$, we also write $\hat{\pi}_i(t_i)$ the *hierarchy of preferences of t_i*, constructed for the induced preference type space $\mathcal{T} = (T_i, \pi_i^{\nu_i, u_i})_{i \in \mathcal{I}}$.

Proposition 2 *For each preference type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$, $(\hat{\pi}_i)_{i \in \mathcal{I}}$ is a preference-preserving mapping from $\mathcal{T}$ to the universal type space $\mathcal{T}^*$.*

We write $\hat{\pi}_i(t_i, \mathcal{T})$ for the hierarchy of preferences of t_i when we emphasize the preference type space $\mathcal{T}$ to which t_i belongs.

Definition 6 *Two types t_i in $\mathcal{T}$ and t_i' in $\mathcal{T}'$ have equivalent preference hierarchies if they map to the same type in T_i^*, i.e., $\hat{\pi}_i(t_i, \mathcal{T}) = \hat{\pi}_i(t_i', \mathcal{T}')$.*

5 Strategic Distinguishability

To give a characterization of equilibrium strategic distinguishability, we must require types to be countable in order to ensure existence. Now we have:

Theorem 1 *Two countable types are strategically indistinguishable if and only if they have equivalent preference hierarchies.*

Countability is required only to show the existence of a local equilibrium, and any other set of conditions ensuring existence of a local equilibrium would be sufficient. Proposition 3 below establishes that if two types have equivalent preference hierarchies, then they are strategically indistinguishable. The argument is as follows: suppose agent i expects other agents to follow strategies that are measurable with respect to their higher-order preferences. Then it is a best response to choose a strategy that is measurable with respect to his own higher-order preferences. To show the converse, we will construct a mechanism in which any pair of types that do not have equivalent preference hierarchies have disjoint equilibrium actions. We postpone this proof to Section 6.2.

Lemma 1 *For every pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$, if $\varphi = (\varphi_i)_{i \in \mathcal{I}}$ is a preference-preserving mapping from $\mathcal{T}$ to $\mathcal{T}'$, then $E_i(t_i, \mathcal{T}, \mathcal{M}) \supseteq E_i(\varphi_i(t_i), \mathcal{T}', \mathcal{M})$ for every $i \in \mathcal{I}$, $t_i \in T_i$ and mechanism $\mathcal{M}$.*

Proof. Pick any local equilibrium $\sigma' = (\sigma_i')$ of $(\mathcal{T}, \mathcal{M})$ associated with preference-closed subspace $\tilde{T}' = \prod_i T_i'$ of $\mathcal{T}'$. Let $\tilde{T}_i = \varphi_i^{-1}\left(\tilde{T}_i'\right)$ and $\sigma_i = \sigma_i' \circ \varphi_i$. Since φ preserves preferences, $\tilde{T} = \prod_i \tilde{T}_i$ is a preference-closed subspace of $\mathcal{T}$ and $\sigma = (\sigma_i)$ is a local equilibrium of $(\mathcal{T}, \mathcal{M})$ associated with $\tilde{T}$. ■

Proposition 3 *For two countable types t_i in $\mathcal{T}$ and t_i' in $\mathcal{T}'$ with $\hat{\pi}_i(t_i, \mathcal{T}) = \hat{\pi}_i(t_i', \mathcal{T}')$, we have $E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i(t_i', \mathcal{T}', \mathcal{M}) \neq \emptyset$ for any mechanism $\mathcal{M}$.*

Proof. By Proposition 2, $\hat{\pi}(\cdot, \mathcal{T})$ and $\hat{\pi}(\cdot, \mathcal{T}')$ are preference-preserving mappings from $\mathcal{T}$ and $\mathcal{T}'$ to the universal space $\mathcal{T}^*$, respectively. By Lemma 1, we have $E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i(t_i', \mathcal{T}', \mathcal{M}) \supseteq E_i(t_i^*, \mathcal{T}^*, \mathcal{M})$, where $t_i^* = \hat{\pi}_i(t_i, \mathcal{T}) = \hat{\pi}_i(t_i', \mathcal{T}')$. Since t_i is countable in $\mathcal{T}$, t_i^* is also countable in $\mathcal{T}^*$, thus $E_i(t_i^*, \mathcal{T}^*, \mathcal{M}) \neq \emptyset$. ■

6 Rationalizability

We introduce a natural definition of rationality - *interim preference correlated rationalizability* (IPCR) - for the worst outcome preference environments studied in this paper. We then show how our characterization of strategic indistinguishability for equilibrium reported in Theorem 1 continues to hold for this definition of rationalizability. As a corollary, the equivalent preference hierarchies characterize strategic indistinguishability for any solution concept which coarsens equilibrium and refines IPCR. We then report a proof of this result, which will imply the part of Theorem 1 which we did not yet prove.

6.1 Interim Preference Correlated Rationalizability

Fix a preference type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$. Write $\Gamma_i \colon T_i \rightrightarrows A_i$ for a correspondence specifying for each type t_i of agent i, a set of actions $\Gamma_i(t_i)$ that are available to type t_i. Fix a profile Γ_{-i} of correspondences of all agents except i. Suppose that agent i were convinced that each agent j of type t_j will choose an action in $\Gamma_j(t_j)$. We will say that action a_i is a best response for t_i against Γ_{-i} if there exists a preference for type t_i in $P_i(\Theta \times T_{-i} \times A_{-i})$ under which (1) there is certainty that action-type profiles of agents other than i are consistent with Γ_{-i}; (2) the marginal preference over $F(\Theta \times T_{-i})$ is consistent with type t_i's original preferences; and (3) a_i is a best response. A correspondence profile $\Gamma = (\Gamma_i)_{i \in \mathcal{I}}$ is a best response correspondence if every action allowed for any type of any agent is a best response to the behavior of other agents. An action is interim preference correlated rationalizable for a given type if it is a possible action for that type in a best response correspondence. More formally:

Definition 7 *Fix a type space $\mathcal{T}$ and a mechanism $\mathcal{M}$. An action $a_i \in A_i$ is a best reply for type $t_i \in T_i$ against Γ_{-i} if there exists $\succsim_i \in P_i(\Theta \times T_{-i} \times A_{-i})$ such that $\Theta \times \text{graph}(\Gamma_{-i})$ is $\succsim_i$-certain, $\text{mrg}_{\Theta \times T_{-i}} \succsim_i = \pi_i(t_i)$ and*

$$\forall a_i' \in A_i, \quad g(\cdot, a_i, \cdot) \; (\text{mrg}_{\Theta \times A_{-i}} \succsim_i) \; g(\cdot, a_i', \cdot).$$

$\Gamma = (\Gamma_i)_{i \in \mathcal{I}}$ is a best reply correspondence if, for every $i \in \mathcal{I}$, $t_i \in T_i$, and $a_i \in \Gamma_i(t_i)$, a_i is a best reply for type t_i against Γ_{-i}. An action a_i is interim preference correlated rationalizable (IPCR) for type t_i if there exists a best reply correspondence Γ with $\Gamma_i(t_i) \ni a_i$.

We write $R_i(t_i, \mathcal{T}, \mathcal{M})$ for the set of IPCR actions for type t_i in type space $\mathcal{T}$ and mechanism $\mathcal{M}$. As usual, we can define $R_i(t_i, \mathcal{T}, \mathcal{M})$ recursively: let $R_{i,0}(t_i, \mathcal{T}, \mathcal{M}) = A_i$ for every $i \in \mathcal{I}$ and $t_i \in T_i$, and, for every $n \geq 1$, let $R_{i,n}(t_i, \mathcal{T}, \mathcal{M})$ be the set of all best replies for type t_i against $R_{-i,n-1}(\cdot, \mathcal{T}, \mathcal{M})$. One can show that $R_i(t_i, \mathcal{T}, \mathcal{M}) = \bigcap_{n \geq 0} R_{i,n}(t_i, \mathcal{T}, \mathcal{M})$, which is nonempty.

IPCR is a very permissive notion of rationalizability. In particular, it allows agents to believe that others' actions convey information about their own preferences over outcomes (consistent with the maintained worst outcome assumption). In the example of Section 3.2, all actions were IPCR even though `optout` was a dominant action if one assumed that the opponent's action did not convey payoff relevant information. Morris and Takahashi (2011) show a formal sense in which this definition of rationalizability captures the implications of common certainty of rationality, under the assumption of expected utility preferences that respect worst outcomes.

Definition 8 *Two types of agent i, t_i in $\mathcal{T}$ and t_i' in $\mathcal{T}'$, are IPCR strategically indistinguishable if, for every mechanism $\mathcal{M}$, there exists some action that can be chosen by both types, so that $R_i(t_i, \mathcal{T}, \mathcal{M}) \cap R_i(t_i', \mathcal{T}', \mathcal{M}) \neq \emptyset$ for every $\mathcal{M}$. Conversely, t_i and t_i' are IPCR strategically distinguishable if there exists a mechanism in which no action can be chosen by both types, so that $R_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_i(t_i', \mathcal{T}', \mathcal{M}^*) = \emptyset$ for some $\mathcal{M}^*$.*

Theorem 2 *Two types are IPCR strategically indistinguishable if and only if they have equivalent preference hierarchies.*

In the next Sub-Section, we prove that IPCR strategically indistinguishable types have the same preference hierarchies. Under the countability assumption, the other direction - showing that if two types have equivalent preference hierarchies, then they are IPCR strategically indistinguishable - follows from Proposition 3, which proved the corresponding step in Theorem 1, as equilibrium actions are a subset of IPCR actions. However, IPCR actions always exist even for uncountable types. In this case, an analogous argument goes through. In particular, the result follows from Theorem 3, which shows that two types with equivalent preference hierarchies are IPCR strategically equivalent.

6.2 Proof of Theorems 1 and 2

Let d_i^* be a metric compatible with the product topology on the universal space $T_i^* \subset \prod_{n=0}^{\infty} P_i(\Theta \times X_{-i,n})$. The remaining direction of Theorems 1 and 2 follows from the next proposition.

Proposition 4 *For every $\varepsilon > 0$, there exists a mechanism $\mathcal{M}^*$ such that*

$$d_i^*(\hat{\pi}_i(t_i, \mathcal{T}), \hat{\pi}_i(t_i', \mathcal{T}')) > \varepsilon \Rightarrow R_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_i(t_i', \mathcal{T}', \mathcal{M}^*) = \emptyset$$

for every pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$, $i \in I$, $t_i \in T_i$, and $t_i' \in T_i'$.

Note that Proposition 4 is stronger than necessary to prove Theorems 1 and 2. In particular, the construction of $\mathcal{M}^*$ depends on ε, but is independent of any details of type spaces $\mathcal{T}$ and $\mathcal{T}'$ or any pair of two types t_i and t_i' that we want to distinguish.

Abreu and Matsushima (1992b) proved such a result for finite type spaces. In the universal belief type space (the space of Mertens-Zamir hierarchies), Dekel, Fudenberg, and Morris (2006, Lemma 4) construct a discretized direct mechanism in which only actions close to truth telling are interim correlated rationalizable. As we discuss below in Section 7.3, their result corresponds to Proposition 4 under the restriction of common certainty of payoffs. Our proof uses a similar mechanism to both papers, with agents essentially reporting their first level (belief or preference) type, their second level type, and so on. Agents can be given individual incentives to report their first level types truthfully and then inductively, if all agents report their kth level types truthfully, each agent can be given an incentive to report his $(k+1)$th level type truthfully by making outcomes contingent on kth level report of other agents. Two complications may potentially destroy the agents' incentives for truth-telling: (i) Outcomes are not necessarily private goods, and in particular the social planner cannot necessarily give a reward to one agent without affecting the other agents' incentives. Especially, an agent's incentives to report her lower-order preferences are affected by how the social planner uses her reports to solicit other agents' higher-order preferences. (ii) As an agent sends less accurate reports about her lower-order preferences, other agents become less willing to report their higher-order preferences accurately. (i) originates the issue, whereas (ii) "multiplies" it.[5] The finiteness assumption allows Abreu and Matsushima (1992b) to deal with both issues by making higher level reports have uniformly lower impact on agents' preferences than lower level reports. Dekel, Fudenberg, and Morris (2006) implicitly assume private goods, removing problem (i). We must carefully exploit our structural assumptions, such as compactness and metrizability of Θ, continuity and monotone continuity of preferences, and existence of the worst outcome, to deal with these issues from the original truth-telling mechanism. The next two subsections are devoted to the proof of Proposition 4.

6.2.1 Single-Agent Revelation Mechanisms

As a preliminary step, here we analyze a single-agent mechanism that reveals her preferences. In this subsection, fix a compact metric space X of states with metric d. Let d_P be a metric compatible with the topology on $P_w(X)$. For each $\succsim \in P_w(X)$, we define the indicator function of $\succsim$, $\chi_{\succsim}$, that

[5] Inaccurate reports may occur in Dekel, Fudenberg, and Morris (2006), but they come purely from discretization.

maps pairs of acts $f, f' \in F(X)$ to 0, 1/2, or 1 as follows:

$$\chi_{\succsim}(f, f') = \begin{cases} 1 & \text{if } f \succ f', \\ 1/2 & \text{if } f \sim f', \\ 0 & \text{if } f \prec f', \end{cases}$$

for any $f, f' \in F(X)$. Let $F_c(X) \subseteq F(X)$ be the set of continuous acts over X. Since X is a compact metric space, by the Stone-Weierstrass theorem, there exists a countable dense subset $F = \{f_1, f_2, \ldots\} \subset F_c(X)$ in the sup norm. Fix such an F.

We consider the following direct mechanism $\mathcal{M}^0 = (P_w(X), g^0)$ for a single agent with action set $P_w(X)$ and outcome function

$$g^0(\cdot, a) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} 2^{-k-l+1} \chi_a(f_k, f_l) f_k \tag{6}$$

for each $a \in P_w(X)$. Under the mechanism $\mathcal{M}^0$, the agent reports her preference. Then the social planner randomly draws a pair of acts from F and assigns the agent with her preferred act according to her reported preference.[6]

In Lemma 2 below, we show that truth telling is a dominant strategy in $\mathcal{M}^0$ for every type. Indeed, by invoking the compactness of X, we show a "robust" version of strategy proofness: in every mechanism close to $\mathcal{M}^0$, the agent strictly prefers reporting almost true preferences to reporting others according to almost true preferences.

Recall that, for each report $a \in P_w(X)$, $g^0(\cdot, a)$ is an act over X, which determines an outcome z with probability $g^0(x, a)(z)$ when the nature chooses $x \in X$. We consider two sources of perturbations to this act. First,with small probability the outcome may not be chosen according to $g^0(x, a)$. Formally, for each $\delta > 0$ and measurable space C, we consider perturbed outcome function $g \colon X \times P_w(X) \times C \to \Delta(Z)$ such that $|g(\cdot, \cdot, c) - g^0| = \sup_{x \in X, a \in P_w(X)} |g(x, a, c) - g^0(x, a)| \leq \delta$ for every $c \in C$. Second, nature may choose x' in a neighborhood of x when instead nature is supposed to choose x. Formally, for each $\delta > 0$, let D^δ be the δ-neighborhood of the diagonal of $X \times X$, $\{(x, x') \in X \times X \mid d(x, x') \leq \delta\}$. For each $\delta > 0$, $\succsim \in P_w(X)$, and measurable space C, let

$$P_w^{\delta, C}(\succsim) = \left\{ \mathrm{mrg}_{2,3} \succsim' \in P_w(X \times C) \mid \begin{array}{l} \exists \succsim' \in P_w(X \times X \times C) \text{ s.t.:} \\ (1)\ \mathrm{mrg}_1 \succsim' = \succsim, \\ (2)\ D^\delta \times C \text{ is } \succsim'\text{-certain}, \end{array} \right\}, \tag{7}$$

where $\mathrm{mrg}_\Lambda \succsim'$ with $\Lambda \subset \{1, 2, 3\}$ denotes the marginal of $\succsim'$ with respect to the coordinates in Λ. In words, $P_w^{\delta, C}(\succsim)$ is the set of preferences over noisy acts induced by the original preference $\succsim$.

[6] Strictly speaking, $\mathcal{M}^0$ is not a mechanism according to our definition, because its action set is infinite. The mechanism we will construct in the next subsection to prove Proposition 4, however, has finite actions.

Lemma 2 *For every $\varepsilon > 0$, there exists $\delta > 0$ such that the following holds: for every preference $\succsim \in P_w(X)$, every pair of reports $a, b \in P_w(X)$ that satisfy $d_P(\succsim, a) \leq \delta$ and $d_P(\succsim, b) > \varepsilon$, every measurable space C, and every perturbed outcome function $g\colon X \times P_w(X) \times C \to \Delta(Z)$ that satisfies $|g(\cdot,\cdot,c) - g^0| \leq \delta$ for every $c \in C$, the agent strictly prefers $g(\cdot, a, \cdot)$ to $g(\cdot, b, \cdot)$ according to every preference in $P_w^{\delta,C}(\succsim)$.*

Proof. See Appendix. ■

6.2.2 Proof of Proposition 4

Let d_Θ be a metric compatible with the topology on Θ. For each $i \in \mathcal{I}$ and $n \geq 1$, let $d_{P,i,n}$ be a metric compatible with the topology on the set of agent i's n-th order preferences, $P_i(\Theta \times X_{-i,n-1})$, and let $d_{i,n}$ be

$$d_{i,n}((\theta, t_{-i,1}, \ldots, t_{-i,n}), (\theta', t'_{-i,1}, \ldots, t'_{-i,n})) = \max\left\{ d_\Theta(\theta, \theta'), \max_{1 \leq k \leq n, j \neq i} d_{P,j,k}(t_{j,k}, t'_{j,k}) \right\},$$

which is a metric compatible with the product topology on $\Theta \times X_{-i,n} = \Theta \times \prod_{k=0}^{n-1} \prod_{j \neq i} P_j(\Theta \times X_{-j,k})$.

Fix any $\varepsilon > 0$. Recall that d_i^* is a metric compatible with the product topology on $T_i^* \subset \prod_{n=0}^{\infty} P_i(\Theta \times X_{-i,n})$. By the definition of the product topology, there exist $\bar{\varepsilon} > 0$ and $N \in \mathbb{N}$ such that, for every $t_i = \{t_{i,n}\}_{n=1}^{\infty}, t'_i = \{t'_{i,n}\}_{n=1}^{\infty} \in T_i^*$, if $d_i^*(t_i, t'_i) > \varepsilon$, then there exists some $n \leq N$ such that $d_{P,i,n}(t_{i,n}, t'_{i,n}) > \bar{\varepsilon}$. Pick such $\bar{\varepsilon}$ and N.

For each $i \in \mathcal{I}$ and $n \leq N$, substitute $X = \Theta \times X_{-i,n-1}$, $d = d_{i,n-1}$, and $d_P = d_{P,i,n}$ in Section 6.2.1. Pick a countable dense subset of $F_c(\Theta \times X_{-i,n-1})$, and define $g_{i,n}^0\colon \Theta \times X_{-i,n-1} \times P_i(\Theta \times X_{-i,n-1}) \to \Delta(Z)$ as in (6). For $\delta > 0$, define $D_{i,n}^\delta$ as the δ-neighborhood of the diagonal of $\Theta \times X_{-i,n-1} \times \Theta \times X_{-i,n-1}$. For $\delta > 0$, $\succsim_{i,n} \in P_i(\Theta \times X_{-i,n-1})$, and measurable space C, define $P_{i,n}^{\delta,C}(\succsim_{i,n})$ as in (7). By Lemma 2, there exist $0 < \varepsilon_0 \leq \varepsilon_1 \leq \cdots \leq \varepsilon_{N-1} \leq \varepsilon_N \leq \bar{\varepsilon}/2$ such that, for every $i \in \mathcal{I}$ and $n \leq N$, for every preference $\succsim_{i,n} \in P_i(\Theta \times X_{-i,n-1})$, every pair of reports $a_{i,n}, b_{i,n} \in P_i(\Theta \times X_{-i,n-1})$ that satisfy $d_{P,i,n}(\succsim_{i,n}, a_{i,n}) \leq \varepsilon_{n-1}$ and $d_{P,i,n}(\succsim_{i,n}, b_{i,n}) > \varepsilon_n$, every measurable space C, and every perturbed outcome function $g_{i,n}\colon \Theta \times X_{-i,n-1} \times P_i(\Theta \times X_{-i,n-1}) \times C \to \Delta(Z)$ that satisfies $|g_{i,n}(\cdot,\cdot,c) - g_{i,n}^0| \leq \varepsilon_{n-1}$ for every $c \in C$, agent i strictly prefers $g_{i,n}(\cdot, a_{i,n}, \cdot)$ to $g_{i,n}(\cdot, b_{i,n}, \cdot)$ according to every preference in $P_{i,n}^{\varepsilon_{n-1},C}(\succsim_{i,n})$.

We define a mechanism $\mathcal{M}^* = ((A_i^*)_{i \in \mathcal{I}}, g^*)$ as follows. For each $i \in \mathcal{I}$ and $n \leq N$, let $A_{i,n}^*$ be any ε_{n-1}-dense finite subset of $P_i(\Theta \times X_{-i,n-1})$ with respect to $d_{P,i,n}$, and $A_i^* = \prod_{n=1}^{N} A_{i,n}^*$. Define $g^*\colon \Theta \times A^* \to \Delta(Z)$ by

$$g^*(\theta, a) = \frac{1-\delta}{I(1-\delta^N)} \sum_{i=1}^{I} \sum_{n=1}^{N} \delta^{n-1} g_{i,n}^0(\theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{i,n})$$

for each $\theta \in \Theta$ and $a = (a_{i,n}) \in A^*$, where $\delta > 0$ is small enough to satisfy $(1-\delta)/\delta \geq (I-1)(1-\varepsilon_0)/\varepsilon_0$.

Claim 1 *For every type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ and $n \leq N$, we have*

$$a_i \in R_{i,n}(t_i, \mathcal{T}, \mathcal{M}^*) \Rightarrow d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), a_{i,n}) \leq \varepsilon_n$$

for every $i \in \mathcal{I}$ and $t_i \in T_i$.

Proof. The proof is by induction on n. Suppose that, for every $k \leq n-1$, $a_i \in R_{i,n-1}(t_i, \mathcal{T}, \mathcal{M}^*)$ implies $d_{P,i,k}(\hat{\pi}_{i,k}(t_i, \mathcal{T}), a_{i,k}) \leq \varepsilon_k \leq \varepsilon_{n-1}$ for every $i \in \mathcal{I}$ and $t_i \in T_i$. Suppose that there exists $a_i^* \in R_{i,n}(t_i, \mathcal{T}, \mathcal{M}^*)$ such that $d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), a_{i,n}^*) > \varepsilon_n$. Then there exists $\succsim_i \in P_i(\Theta \times T_{-i} \times A_{-i}^*)$ such that $\Theta \times \mathrm{graph}(R_{-i,n-1}(\cdot, \mathcal{T}, \mathcal{M}^*))$ is $\succsim_i$-certain, $\mathrm{mrg}_{\Theta \times T_{-i}} \succsim_i = \pi_i(t_i)$, and agent i weakly prefers $g^*(\cdot, a_i^*, \cdot)$ to $g^*(\cdot, a_i, \cdot)$ for every $a_i \in A_i^*$ according to $\mathrm{mrg}_{\Theta \times A_{-i}^*} \succsim_i$.

Let $C = \prod_{k=n}^N A_{-i,k}^*$ and $\varphi_{-i} \colon \Theta \times T_{-i} \times A_{-i}^* \to \Theta \times X_{-i,n-1} \times \Theta \times X_{-i,n-1} \times C$ such that $\varphi_{-i}(\theta, t_{-i}, a_{-i}) = (\theta, \hat{\pi}_{-i,1}(t_{-i}, \mathcal{T}), \ldots, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), \theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{-i,n}, \ldots, a_{-i,N})$. Collect all the terms in g^* that depend on $a_{i,n}$ and define $g_{i,n}^* \colon \Theta \times X_{-i,n-1} \times A_{i,n}^* \times C \to \Delta(Z)$ by

$$\begin{aligned} &g_{i,n}^*(\theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{i,n}, a_{-i,n}, \ldots, a_{-i,N}) \\ &= K \left(g_{i,n}^0(\theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{i,n}) + \sum_{j \in \mathcal{I} \setminus \{i\}} \sum_{k=n+1}^N \delta^{k-n} g_{j,k}^0(\theta, a_{-j,1}, \ldots, a_{-j,k-1}, a_{j,k}) \right), \end{aligned}$$

where $a_{i,k} = a_{i,k}^*$ for $k \neq n$ when they appear in the second term, and K is a positive normalization constant. Since we chose sufficiently small δ, we have $|g_{i,n}^*(\cdot, \cdot, c) - g_{i,n}^0| \leq \varepsilon_0 \leq \varepsilon_{n-1}$ for every $c \in C$. Let $\succsim_i' = (\varphi_{-i})^P(\succsim_i)$. By the induction hypothesis, $\varphi_{-i}(\Theta \times \mathrm{graph}(R_{-i,n-1}(\cdot, \mathcal{T}, \mathcal{M}^*))) \subseteq D_{i,n}^{\varepsilon_{n-1}} \times C$ is $\succsim_i'$-certain. Thus, we have $\mathrm{mrg}_{\Theta \times A_{-i}^*} \succsim_i \in P_{i,n}^{\varepsilon_{n-1}, C}(\hat{\pi}_{i,n}(t_i, \mathcal{T}))$. Since $A_{i,n}^*$ is ε_{n-1}-dense in $P_i(\Theta \times X_{-i,n-1})$, there exists $a_{i,n}' \in A_{i,n}^*$ such that $d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), a_{i,n}') \leq \varepsilon_{n-1}$. By Lemma 2, $\mathrm{mrg}_{\Theta \times A_{-i}^*} \succsim_i$ strictly prefers $g_{i,n}^*(\cdot, a_{i,n}', \cdot)$ to $g_{i,n}^*(\cdot, a_{i,n}^*, \cdot)$, thus $\mathrm{mrg}_{\Theta \times A_{-i}^*} \succsim_i$ strictly prefers $g^*(\cdot, a_{i,n}', a_{i,-n}^*, \cdot)$ to $g^*(\cdot, a_i^*, \cdot)$. This is a contradiction. ■

We can now complete the proof of Proposition 4.

Proof of Proposition 4. Pick any pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$, $i \in \mathcal{I}$, $t_i \in T_i$, and $t_i' \in T_i'$. Suppose that there exists $a_i = (a_{i,1}, \ldots, a_{i,N}) \in R_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_i(t_i', \mathcal{T}', \mathcal{M}^*)$. For every $n \leq N$, since $a_i \in R_{i,n}(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_{i,n}(t_i', \mathcal{T}', \mathcal{M}^*)$, we have

$$\begin{aligned} &d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), \hat{\pi}_{i,n}(t_i', \mathcal{T}')) \\ &\leq d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), a_{i,n}) + d_{P,i,n}(\hat{\pi}_{i,n}(t_i', \mathcal{T}'), a_{i,n}) \leq 2\varepsilon_n \leq \bar{\varepsilon} \end{aligned}$$

by Claim 1. Thus $d_i^*(\hat{\pi}_i(t_i, \mathcal{T}), \hat{\pi}_i(t_i', \mathcal{T}')) \leq \varepsilon$. ■

7 Rationalizability and Strategic Equivalence

Our notion of strategic distinguishability is very demanding: in some game, two types have no equilibrium (or rationalizable) actions in common. The notion of strategic indistinguishability is correspondingly undemanding: it is enough that the two types have some equilibrium (or rationalizable) action in common in every game. In this section, we will study the alternative notion of strategic equivalence. Two types are strategically equivalent if they have the same set of equilibrium (or rationalizable) actions. For any (nonempty-valued) solution concept, strategic equivalence is a stronger requirement than strategic indistinguishability and thus implies a finer partition of types. The corresponding notion of strategic non-equivalence will then be easier to satisfy than strategic distinguishability.

While the characterization of strategic distinguishability is the same for most solution concepts (i.e., for equilibrium, interim preference correlated rationalizability, and, we will show, everything in between), we will see that strategic equivalence characterizations are sensitive to the solution concept. To understand strategic equivalence and its sensitivity, it is useful to introduce a family of rationalizability notions refining interim preference correlated rationalizability, which impose restrictions on the preferences supporting a best response. Our definition of IPCR allows agents' ex post preferences over lotteries, conditional on others' actions and types, to be anything consistent with the worst outcome assumption. Suppose that we impose a further restriction on agents' possible ex post preferences. A given restriction then gives rise to a definition of rationalizability, where preferences supporting a best response must have ex post preferences consistent with the restriction. We show that if we restrict attention to types that belong to type spaces where a given preference restriction holds, then two types are strategically equivalent under the version of rationalizability satisfying that restriction if and only if they have equivalent preference hierarchies.

This result has two important special cases. First, if no restrictions other than the worst outcome assumption are imposed on rationalizability, i.e., if we stick to our earlier definition of IPCR, then this result implies that two types are IPCR strategically equivalent if and only if they have equivalent preference hierarchies. Second, if we impose the restriction that ex post preferences are fixed, i.e., there is common certainty of payoffs, then this result reduces to (a generalization of) the result of Dekel, Fudenberg and Morris (2006, 2007) that, with common certainty of payoffs as a maintained assumption, two types have the same interim correlated rationalizable actions if and only if they have the same Mertens-Zamir higher-order belief hierarchy.

7.1 Ex Post Preference Restrictions

For each $\succsim \in P_w(X)$ and measurable $E \subseteq X$, we write $\succsim_E$ for the *conditional preference* over lotteries defined by

$$y \succsim_E y' \Leftrightarrow y_E y'' \succsim y'_E y''$$

for any $y, y' \in \Delta(Z)$ and some $y'' \in \Delta(Z)$. By independence of $\succsim$, the choice of y'' does not affect the definition of $\succsim_E$.

An *ex post restriction* on agents' preferences will specify a set of possible conditional preferences for each agent. Thus $\mathbf{U} = (U_i)_{i \in I}$, where each U_i is a non-empty set of linearly independent vectors in $\Delta(Z \setminus \{w_i\}) \subset \mathbb{R}^{Z \setminus \{w_i\}}$.[7] The interpretation is that we will impose the requirement that agent i's preferences are representable by convex combinations of U_i, even if they are conditioned on observable states and other agents' types and actions.

We will say that agent i's preference relation $\succsim_i \in P_i(X)$ is U_i-*consistent* if, for any non-$\succsim_i$-null event $E \subseteq X$, the conditional preference $\succsim_{i,E}$ is represented by a convex combination of U_i. A type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ is $\mathbf{U}$-*consistent* if, for each $i \in \mathcal{I}$ and $t_i \in T_i$, $\pi_i(t_i)$ is U_i-consistent. A type t_i is $\mathbf{U}$-*consistent* if it belongs to a $\mathbf{U}$-consistent preference-closed subspace.

We can now define a family of rationalizability concepts for a game $(\mathcal{T}, \mathcal{M})$ with a variety of ex post preference restrictions.

Definition 9 *Fix a type space $\mathcal{T}$ and a mechanism $\mathcal{M}$. An action $a_i \in A_i$ is a U_i-best reply for type $t_i \in T_i$ against Γ_{-i} if there exists $\succsim_i \in P_i(\Theta \times T_{-i} \times A_{-i})$ such that $\succsim_i$ is U_i-consistent, $\Theta \times \text{graph}(\Gamma_{-i})$ is $\succsim_i$-certain, $\text{mrg}_{\Theta \times T_{-i}} \succsim_i = \pi_i(t_i)$ and*

$$\forall a'_i \in A_i, \quad g(\cdot, a_i, \cdot) \; (\text{mrg}_{\Theta \times A_{-i}} \succsim_i) \; g(\cdot, a'_i, \cdot).$$

$\Gamma = (\Gamma_i)_{i \in \mathcal{I}}$ is a $\mathbf{U}$-best reply correspondence if, for every $i \in \mathcal{I}$, $t_i \in T_i$, and $a_i \in \Gamma_i(t_i)$, a_i is a U_i-best reply for type t_i against Γ_{-i}. An action a_i is interim $\mathbf{U}$-rationalizable for type t_i if there exists a $\mathbf{U}$-best reply correspondence Γ with $\Gamma_i(t_i) \ni a_i$.

Let $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$ be the set of $\mathbf{U}$-rationalizable actions for type t_i in game $(\mathcal{T}, \mathcal{M})$. Let $R_{i,0}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) = A_i$ for every $i \in \mathcal{I}$ and $t_i \in T_i$, and, for every $n \geq 1$, let $R_{i,n}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$ be the set of U_i-best replies for type t_i against $R_{-i,n-1}^{\mathbf{U}}(\cdot, \mathcal{T}, \mathcal{M})$. We have $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) = \bigcap_{n \geq 0} R_{i,n}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$. Note that $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$ is non-empty if and only if t_i is $\mathbf{U}$-consistent.

[7] Linear independence is a condition imposed on utility representations, but, given the isomorphism between $P_i(\{*\})$ and $\Delta(Z \setminus \{w_i\})$, one can provide an equivalent condition on preferences over lotteries. For more details, see Morris and Takahashi (2011).

Battigalli and Siniscalchi (2003) define a family of definitions of rationalizability, called "Δ-rationalizability", by imposing restrictions on first order beliefs within the solution concept. "Payoffs" are not incorporated in their type spaces and thus they implicitly maintain common certainty of payoffs over outcomes. **U**-rationalizability parallels Δ-rationalizability in imposing restrictions within the solution concept on beliefs/preferences, but these restrictions concern conditional preferences rather than interim beliefs.

We are most interested in two notions of rationalizability, which correspond to the minimal and maximal conditional preference restrictions, respectively. For the minimal case, we have $U_i = \{\bar{u}_i\}$, a singleton, for each agent i. The solution concept $R^{\mathbf{U}}$ then corresponds to "interim correlated rationalizability" with the restriction that agent i's preferences over lotteries are always represented by $\bar{u}_i$. We will discuss this case in detail in Section 7.3. For the maximal case, we have $U_i = \{u_{i,z} \mid z \in Z \setminus \{w_i\}\}$, where $u_{i,z}$ is the unit vector with 1 on outcome z, thus the convex hull of U_i is equal to $\Delta(Z \setminus \{w_i\})$. Then conditional preference restrictions become vacuous, and interim **U**-rationalizability corresponds to IPCR.

Definition 10 *Two types of agent i, t_i in $\mathcal{T}$ and t'_i in $\mathcal{T}'$, are $R^{\mathbf{U}}$ strategically indistinguishable if, for every mechanism $\mathcal{M}$, there exists some action that can be chosen by both types, so that $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) \cap R_i^{\mathbf{U}}(t'_i, \mathcal{T}', \mathcal{M}) \neq \emptyset$ for every $\mathcal{M}$. Conversely, t_i and t'_i are $R^{\mathbf{U}}$ strategically distinguishable if there exists a mechanism in which no action can be chosen by both types, so that $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_i^{\mathbf{U}}(t'_i, \mathcal{T}', \mathcal{M}^*) = \emptyset$ for some $\mathcal{M}^*$.*

An immediate corollary of Theorems 1 and 2 is:

Corollary 1 *For any conditional preference restrictions **U**, two **U**-consistent types are $R^{\mathbf{U}}$ strategically indistinguishable if and only if they have equivalent preference hierarchies.*

7.2 Strategic Equivalence

We informally introduced the notion of strategic equivalence in Section 3.2. A formal definition is as follows:

Definition 11 *Two types of agent i, t_i in $\mathcal{T}$ and t'_i in $\mathcal{T}'$, are $R^{\mathbf{U}}$ strategically equivalent if, for every mechanism $\mathcal{M}$, $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) = R_i^{\mathbf{U}}(t'_i, \mathcal{T}', \mathcal{M})$ for every $\mathcal{M}$.*

Now we have:

Theorem 3 *For any conditional preference restrictions **U**, two **U**-consistent types are $R^{\mathbf{U}}$ strategically equivalent if and only if they have equivalent preference hierarchies.*

We report a proof for finite type spaces. The proof is close to the proof of Proposition 1 of Dekel, Fudenberg and Morris (2007) and the proof for general type spaces mirrors the proof of Lemma 1 of Dekel, Fudenberg and Morris (2007), the extension of Proposition 1 to general type spaces.

Proof. We will establish by induction on $n \geq 1$ that, if $\hat{\pi}_{i,n}(t_i, \mathcal{T}) = \hat{\pi}_{i,n}(t_i', \mathcal{T}')$, then $R^{\mathbf{U}}_{i,n}(t_i, \mathcal{T}, \mathcal{M}) = R^{\mathbf{U}}_{i,n}(t_i', \mathcal{T}', \mathcal{M})$. Suppose that this holds for $n-1$, that $\hat{\pi}_{i,n}(t_i, \mathcal{T}) = \hat{\pi}_{i,n}(t_i', \mathcal{T}')$ and that $a_i \in R^{\mathbf{U}}_{i,n}(t_i, \mathcal{T}, \mathcal{M})$. Let $\mu_i \in \Delta(\Theta \times T_{-i} \times U_i)$ and $\mu_i' \in \Delta\left(\Theta \times T'_{-i} \times U_i\right)$ be probability measures that represent $\pi_i(t_i)$ and $\pi_i'(t_i')$, respectively. Since a_i is a U_i-best reply for t_i against $R^{\mathbf{U}}_{-i,n-1}(\cdot, \mathcal{T}, \mathcal{M})$ in $(\mathcal{T}, \mathcal{M})$, there exists $\nu_i \in \Delta(\Theta \times T_{-i} \times A_{-i} \times U_i)$ such that:

$$(1) \quad \nu_i(\theta, t_{-i}, a_{-i}, u_i) > 0 \Rightarrow a_{-i} \in R^{\mathbf{U}}_{-i,n-1}(t_{-i}, \mathcal{T}, \mathcal{M}),$$

$$(2) \quad \sum_{a_{-i} \in A_{-i}} \nu_i(\theta, t_{-i}, a_{-i}, u_i) = \mu_i(\theta, t_{-i}, u_i) \text{ for all } \theta \in \Theta, t_{-i} \in T_{-i}, u_i \in U_i,$$

$$(3) \quad a_i \in \arg\max_{a_i' \in A_i} \sum_{\theta, t_{-i}, a_{-i}, u_i, z} g(\theta, (a_i', a_{-i}))(z) \nu_i(\theta, t_{-i}, a_{-i}, u_i) u_i(z).$$

Let

$$D_{-i,n-1} = \{\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) \mid t_{-i} \in T_{-i}\}.$$

For $\hat{\pi}_{-i,n-1} \in D_{-i,n-1}$, let

$$\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i) = \sum_{\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) = \hat{\pi}_{-i,n-1}} \mu_i(\theta, t_{-i}, u_i).$$

Since $\hat{\pi}_{i,n}(t_i, \mathcal{T}) = \hat{\pi}_{i,n}(t_i', \mathcal{T}')$, μ_i and μ_i' represent the same n-th order preference. Since U_i is linearly independent, $\mu_i = \mu_i'$ induce the same probability distribution over $\Theta \times D_{-i,n-1} \times U_i$, i.e.,

$$\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i) = \sum_{\hat{\pi}_{-i,n-1}\left(t'_{-i}, \mathcal{T}'\right) = \hat{\pi}_{-i,n-1}} \mu_i'\left(\theta, t'_{-i}, u_i\right)$$

for all $\theta \in \Theta$, $\hat{\pi}_{-i,n-1} \in D_{-i,n-1}$ and $u_i \in U_i$.

For each $(\theta, \hat{\pi}_{-i,n-1}, u_i)$ such that $\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i) > 0$, set

$$\sigma_{-i}(a_{-i} | \theta, \hat{\pi}_{-i,n-1}, u_i) = \frac{1}{\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i)} \sum_{\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) = \hat{\pi}_{-i,n-1}} \nu_i(\theta, t_{-i}, a_{-i}, u_i).$$

Note that, for each (θ, t_{-i}, u_i) such that $\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i) > 0$, we have $\sigma_{-i}(a_{-i} | \theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i) > 0$ only if $a_{-i} \in R^{\mathbf{U}}_{-i,n-1}(t_{-i}, \mathcal{T}, \mathcal{M})$.

Let

$$\nu_i'\left(\theta, t'_{-i}, a_{-i}, u_i\right) = \mu_i'(\theta, t'_{-i}, u_i) \sigma_{-i}\left(a_{-i} | \theta, \hat{\pi}_{-i,n-1}(t'_{-i}, \mathcal{T}'), u_i\right).$$

Note that ν_i' is well defined because, whenever $\mu_i'(\theta, t_{-i}', u_i) > 0$, we have $\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}(t_{-i}', \mathcal{T}'), u_i) > 0$.

Now we show that a_i is a U_i-best reply for t_i against $R_{-i,n-1}(\cdot, \mathcal{T}', \mathcal{M})$ in $(\mathcal{T}', \mathcal{M})$. First, suppose that $\nu_i'\left(\theta, t_{-i}', a_{-i}, u_i\right) > 0$. Then there exists $t_{-i} \in T_{-i}$ such that $\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) = \hat{\pi}_{-i,n-1}(t_{-i}', \mathcal{T}')$. Since we have $\hat{\mu}_i\left(\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i\right) = \hat{\mu}_i\left(\theta, \hat{\pi}_{-i,n-1}(t_{-i}', \mathcal{T}'), u_i\right) > 0$ and $\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i\right) = \sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t_{-i}', \mathcal{T}'), u_i\right) > 0$, we have $a_{-i} \in R_{-i,n-1}^{\mathbf{U}}(t_{-i}, \mathcal{T}, \mathcal{M})$, which is equal to $R_{-i,n-1}^{\mathbf{U}}(t_{-i}', \mathcal{T}', \mathcal{M})$ by the induction hypothesis.

Second, by the construction of ν_i', the marginal distribution of ν_i' over $\Theta \times T_{-i} \times U_i$ is equal to μ_i', which represents $\pi_i'(t_i')$.

Third, since we have

$$\begin{aligned}
\sum_{t_{-i}'} \nu_i'\left(\theta, t_{-i}', a_{-i}, u_i\right) &= \sum_{t_{-i}'} \mu_i'(\theta, t_{-i}', u_i)\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t_{-i}', \mathcal{T}'), u_i\right) \\
&= \sum_{\hat{\pi}_{-i,n-1} \in D_{-i,n-1}} \hat{\mu}_i\left(\theta, \hat{\pi}_{-i,n-1}, u_i\right)\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}, u_i\right) \\
&= \sum_{t_{-i}} \mu_i(\theta, t_{-i}, u_i)\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i\right) \\
&= \sum_{t_{-i}} \nu_i\left(\theta, t_{-i}, a_{-i}, u_i\right),
\end{aligned}$$

ν_i and ν_i' have the same marginal distribution over $\Theta \times A_{-i} \times U_i$. Thus a_i is a best reply with respect to ν_i' in $(\mathcal{T}', \mathcal{M})$. ■

Since IPCR corresponds to vacuous conditional preference restrictions, an immediate corollary is:

Corollary 2 *Two types are IPCR strategically equivalent if and only if they have equivalent preference hierarchies.*

7.3 Common Certainty of "Payoffs"

Dekel, Fudenberg and Morris (2006, 2007) show a strategic equivalence result for the solution concept of ICR. In particular, they consider "games" $\mathcal{G} = \left((A_i)_{i \in \mathcal{I}}, \hat{g}\right)$, where A_i is a finite action set for agent i, and a measurable function $\hat{g}\colon \Theta \times A \to [0,1]^I$ describes "payoffs" as a function of observable states Θ and action profiles. "Payoffs" in their setting correspond to von Neumann-Morgenstern indices in our setting, and since the function $\widehat{g}$ is taken to be common certainty among the agents, it is implicitly assumed that there is common certainty of "payoffs" or von Neumann-Morgenstern indices. DFM show that two types have the same set of interim correlated rationalizable actions in all games $\mathcal{G}$ if and only if they have the same MZ hierarchy of beliefs and

higher-order beliefs about Θ. In particular, Lemma 4 of DFM (2006) establishes that types with distinct MZ hierarchies must have distinct ICR actions; Proposition 1 (for finite type spaces) and Lemma 1 (for infinite type spaces) of DFM (2007) establish that types with the same MZ hierarchy have the same set of ICR actions.

Lemma 4 of DFM (2006) is a special case of our Proposition 4. To see why, let $Z = \prod_i Z_i$ with $Z_i = \{0, 1\}$, and $U_i = \{\overline{u}_i\}$ with $\overline{u}_i(z_1, \ldots, z_I) = z_i$. In this case, any belief type space $\mathcal{T} = (T_i, \mu_i)_{i \in \mathcal{I}}$ with $\mu_i \colon T_i \to \Delta(\Theta \times T_{-i})$ induces a preference type space $\mathcal{T}' = (T_i, \pi_i)_{i \in \mathcal{I}}$ by $\pi_i(t_i) = \succsim^{\mu_i(t_i), \overline{u}_i}$. Then IPCR in $(\mathcal{T}', \mathcal{M})$ is more permissive than **U**-rationalizability in $(\mathcal{T}', \mathcal{M})$, which reduces to ICR in $(\mathcal{T}, \mathcal{G})$ (as defined in DFM (2006, 2007)) in the game $\mathcal{G} = \left((A_i)_{i \in \mathcal{I}}, \hat{g}\right)$ with

$$\hat{g}_i(\theta, a) = \sum_z g(\theta, a)(z)\overline{u}_i(z) = \sum_{z_{-i}} g(\theta, a)(1, z_{-i}).$$

Thus our Proposition 4 implies Lemma 4 of DFM (2006).[8] Similarly, Proposition 1 and Lemma 1 of DFM (2007) are a special case of our Theorem 3.

Examples in DFM (2007) and Ely and Pęski (2006) show that under less permissive versions of rationalizability—for example, IIR in DFM (2007)—MZ types do not characterize strategic equivalence. Ely and Pęski (2006) provide a characterization of strategic equivalence for IIR in two-agent games. Liu (2009) and Sadzik (2010) provide characterizations of "redundant" components required for equilibrium strategic equivalence. Thus the message of this "common certainty of payoffs" literature is that strategic equivalence is sensitive to the solution concept considered. Although the point was not highlighted in this literature, it is easy to see that Mertens-Zamir higher-order beliefs characterize strategic distinguishability in this common certainty of payoffs setting. Our Corollary 1 makes this point without common certainty of payoffs.

Thus there is a clean parallel between results for the two environments of "common certainty of payoffs" literature and the general case studied in this paper. Independent of the solution concept, strategic distinguishability is characterized by MZ higher-order beliefs and higher-order preferences, respectively. Characterizations of strategic equivalence depend on the solution concept. ICR strategic equivalence is characterized by MZ higher-order beliefs, and IPCR strategic equivalence is characterized by higher-order preferences. More refined solution concepts may require finer descriptions of types to characterize strategic equivalence.

[8] Indeed, one can show from our Proposition 4 that MZ hierarchies characterize strategic distinguishability even if restrictions are imposed on payoffs across agents. That is, one can use only measurable functions $\hat{g} \colon \Theta \times A \to V$ to strategically distinguish distinct MZ types, where V is a convex subset of $[0, 1]^I$ such that, for any agent i, there exist $v^i, \tilde{v}^i \in V$ such that $v^i_i \neq \tilde{v}^i_i$.

8 Discussion

8.1 Relaxing the Worst Outcome Property

We have assumed so far that, for each agent i, there is common certainty that an outcome w_i is worse than any other outcome for that agent. There are two roles which the worst outcome assumption plays in our analysis. First, combined with the non-triviality assumption, it rules out the possibility of types that are completely indifferent between all outcomes. Second, it ensures the space $P_w(X)$ of all possible preferences is isomorphic to $\Delta(X \times (Z \setminus \{w\}))$, which is compact and metrizable if X is compact and metrizable. Both results are indispensable for our results. Clearly, every action is rationalizable for a completely indifferent type and thus such a type cannot be strategically distinguished from any other type. Also, we can show that—even after ruling out complete indifference—if the set of all possible preferences is not compact, then not only do technical difficulties arise in the construction of a universal preference type space, but more importantly, it is no longer the case that two types with distinct preference hierarchies can be strategically distinguished. This point is shown by Claim 3 in Morris and Takahashi (2011) and is related to the negative results in Ledyard (1986).

The worst outcome assumption is a convenient way of ruling out complete indifference and guaranteeing compactness of the space of possible preferences. However, weaker assumptions will work as well. For $\lambda \in (0, 1/2]$, we say that a binary relation $\succsim$ over $F(X)$ is λ-*continuous* if there exist two outcomes $z, z' \in Z$ with $z \succ z'$ and, for every $f, f' \in F(X)$, we have

$$(1-\lambda)\, z + \lambda f \succsim (1-\lambda)\, z' + \lambda f'.$$

For a general state-dependent preference, preferences over outcomes may depend on states. λ-continuity requires that the strength of such state dependency be bounded in the sense that, even if an agent receives state-dependent acts with probability λ, it does not alter her preference between state-independent outcomes z and z'.

The notion of λ-continuity is a weak requirement. To see this, note that every binary relation $\succsim$ over $F(X)$ that satisfies completeness, transitivity, independence, continuity and monotone continuity is represented by a finite signed measure μ on $X \times Z$:

$$f \succsim f' \Leftrightarrow \int_{X\times Z} f(x)(z)d\mu(x,z) \geq \int_{X\times Z} f'(x)(z)d\mu(x,z).$$

If a preference is not indifferent over lotteries, then it is λ-continuous for a sufficiently small $\lambda > 0$. For example, one can take $\lambda > 0$ such that

$$\frac{\lambda}{1-\lambda} \leq \frac{\|\mathrm{mrg}_Z \mu\|}{\|\mu\|},$$

where $\text{mrg}_Z\mu$ is the marginal of μ on Z given by $(\text{mrg}_Z\mu)(\{z\}) := \mu(X\times\{z\})$, and, for $\nu = \mu, \text{mrg}_Z\mu$, $\|\nu\| := \sup_{E,E'}(\nu(E) - \nu(E'))$, where E and E' vary over all measurable sets, denotes the total variation of ν. Also, every preference in $P_w(X)$ is λ-continuous with any $0 < \lambda \leq 1/|Z|$.

Then we focus on preference type spaces where there is common certainty that all agents' preferences are λ-continuous for some fixed $\lambda > 0$. Such spaces include preference type spaces with the worst outcome property, studied in this paper, and other settings, such as finite type spaces of Abreu and Matsushima (1992) and "compact and continuous" type spaces (see Proposition 6 in Bergemann, Morris and Takahashi (2010)).

For such preference type spaces, we can construct a universal preference type space, consisting of coherent hierarchies of preferences, for each $\lambda > 0$. Also, we can show Theorems 1 and 2, i.e., the universal space characterizes strategic distinguishability for equilibrium, interim preference correlated rationalizability that respects λ-continuity, and everything in between.[9] Details are given in Appendix B of the working paper version of this paper, Bergemann, Morris and Takahashi (2010).

8.2 On the Expected Utility Assumption

Another important assumption maintained throughout the paper is that there is common certainty that all agents have expected utility preferences. Indeed, in the proof of Proposition 4, we used a convex combination of reporting mechanisms to provide a separate incentive, for each agent and each level of his preference hierarchy, to report the preference truthfully, which relies on expected utility preferences, especially on the independence axiom. We conjecture that, if we dropped this assumption and allowed for ambiguity-averse preferences for example, then it would remain true that two types with the same higher order preferences are strategically indistinguishable (under an appropriate definition of higher order preferences), but it is not clear if two types with different attitudes toward ambiguity would be strategically distinguishable. That they might not be is suggested by an impossibility result shown by Chen and Luo (2011), which states that, in the complete information setting with common certainty of "payoffs," if the game is "concave-like," then an agent with a general class of preferences has the same set of rationalizable actions as the agent with expected utility preference.

8.3 Dynamic Mechanisms and Sequential Rationality

Two key findings in our setting are that information and psychological reasons for interdependent preferences cannot be disentangled, and interdependent preferences cannot be observed in a com-

[9] We do not expect to have a strategic equivalence result such as Theorem 3.

plete information setting. For example, suppose that there is common certainty that (i) agent 1 is an altruist (who cares about agent 2's private good consumption) or selfish (caring only about her own consumption); (ii) while agent 2 is either an altruist (who cares about agent 1's private good consumption) or a conditional altruist (caring about agent 1's private good consumption only if agent 1 is an altruist); but (iii) agent 2 is certain that agent 1 is an altruist. Our analysis would say that the two types of agent 2 cannot be strategically distinguished.

We noted in the introduction that this implies that our universal space is much coarser than the important construction of Gul and Pesendorfer (2007) which contains much counterfactual information about interdependent preferences. These findings are consequences of our restriction to static games, and to solution concepts that do not incorporate sequential rationality. Thus in our example, agent 2's type could be strategically distinguished if we looked at sequential equilibria of a dynamic game, like the ultimatum game, by examining if agent 2 was nice to agent 1 after observing the (ex ante zero probability) event that agent 1 was not nice to agent 2. An interesting topic for future work is the extent to which allowing dynamic games with sequential rationality refinements (where behavior will reflect counterfactual information) can reveal the fine information contained in Gul and Pesendorfer (2007) types. For an expected utility setting, such an analysis would lead to a conditional preference universal space analogous to the conditional probability universal space of Battigalli and Siniscalchi (1999). Recent work on dynamic mechanism design in "payoff type" environments, as Müller (2009) and Penta (2009), indirectly addresses these issues.

8.4 Virtual Bayesian Implementation

As we discussed in the introduction, we extend a key lemma in Abreu and Matsushima (1992b) in Proposition 4, where we construct a revealing mechanism for an infinite rather than a finite type space. A small proviso to this statement is that, while we allow infinite type spaces, in our main treatment we impose a worst outcome assumption not used in Abreu and Matsushima (1992b); but as we noted in Section 8.1, we can easily incorporate general finite type spaces in our analysis.

Abreu and Matsushima (1992b) used that lemma to show (for finite type spaces) a necessary "measurability" condition for virtually implementing a social choice function under incomplete information: the social choice function must select the same allocation for any pair of types that are strategically indistinguishable. In our infinite state space setting, we conjecture that the analogous necessary condition would be that the social choice function has to be continuous with respect to the topology on types that generates continuous strategic outcomes. The relevant topology would be the analogue, for our universal preference hierarchy, to the strategic topology on the Mertens-Zamir space introduced by Dekel, Fudenberg and Morris (2006) and further characterized by Chen

et al. (2011).

Abreu and Matsushima (1992b) also adapt arguments from the complete information setting (Abreu and Matsushima (1992a)) to show that the measurability condition is essentially sufficient for virtual robust implementation. We have not considered the extension of this argument to infinite type spaces and thus do not know if a sufficiency result could be proved.

8.5 Payoff Type Environments

Bergemann and Morris (2009) analyze virtual implementation with incomplete information, in a "robust" setting where any beliefs and higher order beliefs about agents' payoff relevant types are possible. This required an analysis of a variant of the strategic distinguishability question in this paper. Consider a "payoff type environment," where there is a finite set Z of outcomes and a finite set of agents, $\mathcal{I} = \{1, ..., I\}$, each with a payoff type φ_i drawn from a finite set Φ_i and a (perhaps interdependent) utility function $\hat{u}_i \colon \Phi \times Z \to \mathbb{R}$. Common certainty of utility functions $(\hat{u}_i)_{i \in \mathcal{I}}$ - and thus agents' ex post preference conditional on the profile of known ϕ - is (implicitly) assumed. Now a type space $\mathcal{T} = (T_i, b_i, \tilde{\varphi}_i)_{i \in \mathcal{I}}$ specifies for each agent i a set of possible types T_i, and mappings $b_i \colon T_i \to \Delta(T_{-i})$ and $\tilde{\varphi}_i \colon T_i \to \Phi_i$ identifying the beliefs and (known) own payoff type of each types. Expressing the strategic distinguishability question of Bergemann and Morris (2009) in the language of this paper, we can identify the set of (say) equilibrium actions $E_i(t_i, \mathcal{T}, \mathcal{M})$ of type t_i from type space $\mathcal{T}$ playing mechanism $\mathcal{M}$. Now say that payoff types φ_i and φ_i' are strategically distinguishable if there exists a mechanism where—whatever their beliefs and higher-order beliefs about other agents' payoff types—they have no action in common; formally, if there exists a mechanism $\mathcal{M}^*$ such that for all t_i in type space $\mathcal{T}$ with $\tilde{\varphi}_i(t_i) = \varphi_i$ and t_i' in type space $\mathcal{T}'$ with $\tilde{\varphi}_i'(t_i') = \varphi_i'$,

$$E_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap E_i\left(t_i', \mathcal{T}', \mathcal{M}^*\right) = \emptyset.$$

Conversely, payoff types φ_i and φ_i' are strategically indistinguishable if, for every mechanism $\mathcal{M}$, there exist types t_i in type space $\mathcal{T}$ with $\tilde{\varphi}_i(t_i) = \varphi_i$ and t_i' in type space $\mathcal{T}'$ with $\tilde{\varphi}_i'(t_i') = \varphi_i'$ such that

$$E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i\left(t_i', \mathcal{T}', \mathcal{M}\right) \neq \emptyset.$$

Bergemann and Morris (2009) present a characterization of strategically indistinguishable payoff types and show that strong interdependence in utilities gives rise to strategic indistinguishability. For example, in a quasi-linear environment where agent i has payoff type $\varphi_i \in [0, 1]$ and his valuation of an object is given by $v_i(\varphi) = \varphi_i + \gamma \sum_{j \neq i} \varphi_j$ for some $\gamma \in \mathbb{R}_+$, two distinct payoff types of any agent are strategically distinguishable if and only if $\gamma \leq \frac{1}{I-1}$.

8.6 Strategic Revealed Preference

Suppose we knew that an agent i would choose a_1 when playing mechanism $\mathcal{M}_1$, a_2 when playing mechanism $\mathcal{M}_2$, and so on. This might be because the agent made these choices in real time (and we knew his/her preferences—and implicitly information—were stable over time), or these might reflect hypothetical choices that the agent would make. If we had a finite data set given by $(a_k, \mathcal{M}_k)_{k=1}^K$, we could ask if there exists a type that could have generated that set of data by rational strategic choice. If we interpret rational strategic choice as choosing according to some solution concept, say, IPCR, i.e., then this "strategic revealed preference" question becomes: is there a type t_i in some type space $\mathcal{T}$ such that $a_k \in R_i(t_i, \mathcal{T}, \mathcal{M}_k)$ for every k?

This is a strategic analogue to the classical revealed preference question of Afriat (1967). In the single person case, without the linear indifference curves generated by expected utility preferences over lotteries, we know that a finite data set is consistent with a rational preference if and only if it satisfies the weak axiom of revealed preference (WARP). To get to our strategic revealed preference question described above, we must first require the outcome space to be a lottery space and impose expected utility preferences, which will require independence as well as WARP in the data. Second, we must translate a choice problem, where an agent picks a most preferred outcome from a set of lotteries, to a strategy setting where many agents make simultaneous (and perhaps interdependent) choices. Our mechanism is a many agent choice problem where outcomes depend not only on an agent's choice but also on others' choices.

Our characterization of strategic distinguishability answers a related but different question. Suppose that all the data that you have observed so far are consistent with an agent being type t_i or type t_i'. Does there exist a mechanism by which one could be sure to distinguish them at the next round? It would be a natural next step to ask how much distinguishing could be done with smaller mechanisms and thus give a characterization of behavioral implications of interdependent preferences in a small set of mechanisms rather than quantify over all mechanisms.

There is a small existing literature developing strategic analogues of classic single agent decision theory. Sprumont (2000) considers static Nash equilibrium in static games, and thus may be the closest to our setting. But the extension from one agent to many agent choice problems is carried out in a very different way. First, he does not consider mixed strategies and does not maintain—as we do—the hypothesis of expected utility preferences. Second, and more importantly, our many agent decision problems (mechanisms) put no structure on the set of choices—there may be arbitrary action sets—but the outcome function may impose restrictions. For example, the outcome resulting from one action profile may be identical to that resulting from another action profile, and we implicitly assume that there is common certainty of this fact. By contrast, Sprumont (2000)

fixes agents' finite action sets and studies choices when there is common certainty that they are restricted to subsets of these actions sets. But he imposes no restrictions on how the outcomes from different action profiles may relate to each other.

A Proof of Lemma 2

Suppose not. Then, there exist $\varepsilon > 0$ such that, for every $n \in \mathbb{N}$, there exist $\succsim_n, a_n, b_n \in P_w(X)$, measurable space C_n, perturbed outcome function $g_n \colon X \times P_w(X) \times C_n \to \Delta(Z)$ with $|g_n(\cdot,\cdot,c) - g^0| \leq 1/n$ for every $c \in C_n$, and $\succsim'_n \in P_w(X \times X \times C_n)$ such that $d_P(\succsim_n, a_n) \leq 1/n$, $d_P(\succsim_n, b_n) \geq \varepsilon$, $\mathrm{mrg}_1 \succsim'_n = \succsim_n$, $D^{1/n} \times C_n$ is $\succsim'_n$-certain, and $\mathrm{mrg}_{2,3} \succsim'_n$ weakly prefers $g_n(\cdot, b_n, \cdot)$ to $g_n(\cdot, a_n, \cdot)$. For each n, let $\nu_n \in \Delta(X \times X \times C_n \times (Z \setminus \{w\}))$ be a probability measure that represents $\succsim'_n$. Note that $\mu_n := \mathrm{mrg}_{1,4} \nu_n$ represents $\succsim_n$, and $\nu_n(D^{1/n} \times C_n \times (Z \setminus \{w\})) = 1$.[10] Since X is a compact metric space, by taking a subsequence if necessary, we can find $\succsim^*, b^* \in P_w(X)$ and $\mu^* \in \Delta(X \times (Z \setminus \{w\}))$ such that $\succsim_n \to \succsim^*$, $b_n \to b^*$, and $\mu_n \to \mu^*$ as $n \to \infty$. Note that $a_n \to \succsim^*$ as $n \to \infty$, $\succsim^* \neq b^*$, and μ^* represents $\succsim^*$.

Claim 2 *For every $k_0 \in \mathbb{N}$, there exists $n_0 \in \mathbb{N}$ such that, for every $n \geq n_0$ and $k, l \leq k_0$, if $\succsim_n$ strictly prefers f_k to f_l, then a_n weakly prefers f_k to f_l.*

Proof. Fix any k_0. Suppose not. Then there exists a pair of $k, l \leq k_0$ and a subsequence of $(\succsim_n, a_n)$ such that $\succsim_n$ strictly prefers f_k to f_l, and a_n strictly prefers f_l to f_k. Since $\succsim_n$ and a_n converge to the same limit, this is a contradiction. ■

Claim 3 *There exist k^*, l^* such that $\succsim^*$ strictly prefers f_{k^*} to f_{l^*} while b^* strictly prefers f_{l^*} to f_{k^*}.*

Proof of Claim 3. Since $\succsim^* \neq b^*$, there exist $f, f' \in F_c(X)$ such that $\succsim^*$ and b^* have different preferences between f and f'. Since $\succsim^*$ and b^* satisfy the continuity, we can assume without loss of generality that $\succsim^*$ strictly prefers f to f' and b^* strictly prefers f' to f. (To see this, suppose, for example, that $\succsim^*$ is indifferent between f and f' while b^* strictly prefers f' to f. Then, replace f by $(1-\lambda)f + \lambda f''$ and f' by $(1-\lambda)f' + \lambda f'''$ for sufficiently small $\lambda > 0$, where $\succsim^*$ strictly prefers f'' to f'''. A similar trick works when $\succsim^*$ strictly prefers f to f' while b^* is indifferent between f to f'.) Since F is dense in $F_c(X)$ in the sup norm, by the continuity of $\succsim^*$ and b^*, we can assume $f, f' \in F$ without loss of generality. ■

Claim 4 *There exists $n_0 \in \mathbb{N}$ such that, for every $n \geq n_0$, b_n strictly prefers f_{l^*} to f_{k^*}.*

Proof of Claim 4. Follows from $b_n \to b^*$ as $n \to \infty$. ■

It follows from Claim 3 that there exists $\eta > 0$ such that

$$7\eta < 2^{-k^*-l^*+1} \int (f_{k^*} - f_{l^*}) d\mu^*.$$

[10] $\mathrm{mrg}_\Lambda \nu_n$ with $\Lambda \subset \{1,2,3,4\}$ denotes the marginal of ν_n with respect to the coordinates in Λ.

Pick $k_0 \geq \max\{k^*, l^*\}$ such that

$$\sum_{\max\{k,l\}>k_0} 2^{-k-l+1} < \eta.$$

Claim 5 *There exists $n_1 \in \mathbb{N}$ such that, for every $n \geq n_1$ and $k, l \in \mathbb{N}$ such that $\max\{k,l\} \leq k_0$, if $\succsim^*$ strictly prefers f_k to f_l, then a_n also strictly prefers f_k to f_l.*

Proof of Claim 5. Follows from $a_n \to \succsim^*$ as $n \to \infty$. ■

Note that

$$\begin{aligned}&(\chi_{a_n}(f_k,f_l) - \chi_{b_n}(f_k,f_l)) \int f_k d\mu^* + (\chi_{a_n}(f_l,f_k) - \chi_{b_n}(f_l,f_k)) \int f_l d\mu^* \\ &= (\chi_{a_n}(f_k,f_l) - \chi_{b_n}(f_k,f_l)) \int (f_k - f_l) d\mu^*\end{aligned}$$

since $\chi_{a_n}(f_l,f_k) = 1 - \chi_{a_n}(f_k,f_l)$ and $\chi_{b_n}(f_k,f_l) = 1 - \chi_{b_n}(f_l,f_k)$.

Claim 6 *For every $n \geq \max\{n_0, n_1\}$, we have*

$$(\chi_{a_n}(f_k,f_l) - \chi_{b_n}(f_k,f_l)) \int (f_k - f_l) d\mu^* \begin{cases} = \int (f_{k^*} - f_{l^*}) d\mu^* & \text{if } (k,l) = (k^*,l^*), \\ \geq 0 & \text{if } \max\{k,l\} \leq k_0. \end{cases}$$

Proof of Claim 6. By Claims 4 and 5, $\chi_{a_n}(f_{k^*}, f_{l^*}) = 1$ and $\chi_{b_n}(f_{k^*}, f_{l^*}) = 0$; $\chi_{a_n}(f_k,f_l) = 1 \geq \chi_{b_n}(f_k,f_l)$ and $\int (f_k - f_l) d\mu^* > 0$ if $\succsim^*$ strictly prefers f_k to f_l; $\chi_{a_n}(f_k,f_l) = 0 \leq \chi_{b_n}(f_k,f_l)$ and $\int (f_k - f_l) d\mu^* < 0$ if $\succsim^*$ strictly prefers f_l to f_k; $\int (f_k - f_l) d\mu^* = 0$ if $\succsim^*$ is indifferent between f_k and f_l. ■

Claim 7 *There exists $n_2 \in \mathbb{N}$ such that, for every $n \geq n_2$ and $k \leq k_0$, we have*

$$\left| \int f_k d(\mathrm{mrg}_{2,4} \nu_n) - \int f_k d\mu_n \right| \leq \eta.$$

Proof of Claim 7. Since X is a compact metric space, every continuous function is uniformly continuous. Therefore, there exists $n_2 \in \mathbb{N}$ such that $|f_k(x) - f_k(x')| \leq \eta$ for every $k \leq k_0$ and $(x,x') \in D^{1/n_2}$. For every $n \geq n_2$, we have

$$\begin{aligned}&\left| \int f_k d(\mathrm{mrg}_{2,4} \nu_n) - \int f_k d\mu_n \right| \\ &= \left| \int (f_k(x')(z) - f_k(x)(z)) d(\mathrm{mrg}_{1,2,4} \nu_n)(x,x',z) \right| \\ &\leq \int |f_k(x')(z) - f_k(x)(z)| d(\mathrm{mrg}_{1,2,4} \nu_n)(x,x',z) \leq \eta\end{aligned}$$

since $|f_k(x')(z) - f_k(x)(z)| \leq \eta$ for $(\mathrm{mrg}_{1,2,4} \nu_n)$-almost every (x,x',z). ■

We can now complete the proof of Lemma 2. Since $\mu_n \to \mu^*$ as $n \to \infty$, there exists $n \geq \max\{n_0, n_1, n_2, 1/\eta\}$ such that, for every $k \leq k_0$, $|\int f_k d\mu_n - \int f_k d\mu^*| < \eta$. We decompose $\int (g_n(\cdot, a_n, \cdot) - g_n(\cdot, b_n, \cdot)) d(\mathrm{mrg}_{2,3,4}\nu_n)$ into the following four terms:

$$\begin{aligned}
&\int (g_n(\cdot, a_n, \cdot) - g_n(\cdot, b_n, \cdot)) d(\mathrm{mrg}_{2,3,4}\nu_n) \\
&= \sum_{\max\{k,l\} \leq k_0} 2^{-k-l+1} (\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)) \int f_k d\mu^* \\
&\quad + \sum_{\max\{k,l\} \leq k_0} 2^{-k-l+1} (\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)) \left(\int f_k d(\mathrm{mrg}_{2,4}\nu_n) - \int f_k d\mu^* \right) \\
&\quad + \sum_{\max\{k,l\} > k_0} 2^{-k-l+1} (\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)) \int f_k d(\mathrm{mrg}_{2,4}\nu_n) \\
&\quad + \int [(g_n(\cdot, a_n, \cdot) - g^0(\cdot, a_n)) - (g_n(\cdot, b_n, \cdot) - g^0(\cdot, b_n))] d(\mathrm{mrg}_{2,3,4}\nu_n).
\end{aligned}$$

The first term is larger than 7η by Claim 6. The other terms are at least as large as -4η, $-\eta$, and -2η, respectively, since $\sum_{\max\{k,l\} \leq k_0} 2^{-k-l+1} < 2$, $|\chi_{a_n} - \chi_{b_n}| \leq 1$,

$$\begin{aligned}
&\left| \int f_k d(\mathrm{mrg}_{2,4}\nu_n) - \int f_k d\mu^* \right| \\
&\leq \left| \int f_k d(\mathrm{mrg}_{2,4}\nu_n) - \int f_k d\mu_n \right| + \left| \int f_k d\mu_n - \int f_k d\mu^* \right| \\
&< 2\eta
\end{aligned}$$

by Claim 7, $\sum_{\max\{k,l\} > k_0} 2^{-k-l+1} < \eta$, $|f_k| \leq 1$, and $|g_n(\cdot, \cdot, c) - g^0| \leq 1/n \leq \eta$ for every $c \in C_n$. Thus $\succsim'_n$ strictly prefers $g_n(\cdot, a_n, \cdot)$ to $g_n(\cdot, b_n, \cdot)$, which is a contradiction.■

References

[1] D. Abreu and H. Matsushima (1992a), "Virtual Implementation in Iteratively Undominated Actions: Complete Information," *Econometrica* 60, 993-1008.

[2] D. Abreu and H. Matsushima (1992b), "Virtual Implementation in Iteratively Undominated Actions: Incomplete Information," at http://www.princeton.edu/~dabreu/index_files/virtual%20implementation-incomplete.pdf

[3] S. Afriat (1967), "The Construction of a Utility Function from Expenditure Data," *International Economic Review* 8, 66–77.

[4] F. J. Anscombe and R. J. Aumann (1963), "A Definition of Subjective Probability," *Annals of Mathematical Statistics* 34, 199–205.

[5] P. Battigalli and M. Siniscalchi (1999), "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games," *Journal of Economic Theory* 88, 188-230.

[6] P. Battigalli and M. Siniscalchi (2003), "Rationalization and Incomplete Information," *Advances in Theoretical Economics* 3 (1), Article 3.

[7] D. Bergemann and S. Morris (2009), "Virtual Robust Implementation," *Theoretical Economics* 4, 45–88.

[8] D. Bergemann, S. Morris and S. Takahashi (2010), "Interdependent Preferences and Strategic Distinguishability," Princeton University Economic Theory Center Working Paper 10-08, http://ssrn.com/abstract=1729280.

[9] A. Brandenburger and E. Dekel (1993), "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory* 59, 189–198.

[10] Y. Chen and X. Luo (2011), "An Indistinguishability Result on Rationalizability Under General Preferences," forthcoming in *Economic Theory*.

[11] Y. Chen, E. Faingold, A. Di Tillio, and S. Yang (2011), "The Strategic Impact of Higher-Order Beliefs.".

[12] E. Dekel, D. Fudenberg, and S. Morris (2006), "Topologies on Types," *Theoretical Economics* 1, 275–309.

[13] E. Dekel, D. Fudenberg, and S. Morris (2007), "Interim Correlated Rationalizability," *Theoretical Economics* 2, 15–40.

[14] A. Di Tillio (2008), "Subjective Expected Utility in Games," *Theoretical Economics* 3, 287–323.

[15] J. C. Ely and M. Pęski (2006), "Hierarchies of Belief and Interim Rationalizability," *Theoretical Economics* 1, 19–65.

[16] L. G. Epstein and T. Wang (1996), ""Beliefs about Beliefs" without Probabilities," *Econometrica* 64, 1343–1373.

[17] F. Gul and W. Pesendorfer (2007), "The Canonical Space for Behavioral Types."

[18] J. O. Ledyard (1986), "The Scope of the Hypothesis of Bayesian Equilibrium," *Journal of Economic Theory* 39 (1), 59–82.

[19] D. K. Levine (1998), "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics* 1, 593-622.

[20] Q. Liu (2009), "On Redundant Types and Bayesian Formulation of Incomplete Information," *Journal of Economic Theory* 144, 2115-2145.

[21] J.-F. Mertens and S. Zamir (1985), "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory* 14 (1), 1–29.

[22] P. Milgrom (2004), *Putting Auction Theory to Work*. Cambridge, England: Cambridge University Press.

[23] S. Morris and S. Takahashi (2011), "Common Certainty of Rationality Revisited."

[24] C. Müller (2009), "Robust Virtual Implementation under Common Strong Belief in Rationality."

[25] R. Myerson (1991), *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

[26] H. Paarsch and H. Hong (2006), *An Introduction to the Structural Econometrics of Auction Data*. Cambridge, MA: M.I.T. Press.

[27] A. Penta (2009), "Robust Dynamic Mechanism Design."

[28] T. Sadzik (2010), "Beliefs Revealed in Bayesian-Nash Equilibrium."

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

[30] J. Weibull (2004), "Testing Game Theory," in *Advances in Understanding Strategic Behavior; Game Theory, Experiments and Bounded Rationality. Essays in Honour of Werner Güth.* Edited by Steffen Huck. Hampshire: Palgrave Macmillan, 85-104.