INTERDEPENDENT PREFERENCES AND STRATEGIC DISTINGUISHABILITY

By

Dirk Bergemann, Stephen Morris and Satoru Takahashi

September 2010
Revised July 2014

COWLES FOUNDATION DISCUSSION PAPER NO. 1772RR

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§

First Version: October 2010
This Version: July 2014

Abstract

We identify a universal type space of possible interdependent (expected utility) preferences of a group of agents satisfying two criteria. First, a type consists of a "detail free" description, in a natural language, of the agents' interdependent preferences. Second, distinct types in the universal type space must be "strategically distinguishable" in the sense that there must exist a mechanism where those types are guaranteed to behave differently in equilibrium.

Our results generalize and unify results of Abreu and Matsushima (1992b) (who characterized strategic distinguishability on fixed finite type spaces) and Dekel, Fudenberg, and Morris (2006), (2007) (who characterized strategic distinguishability on type spaces without preference uncertainty and thus without interdependent preferences).

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 Grants SES 0851200 and 1215808. 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

§National University of Singapore, ecsst@nus.edu.sg

1 Introduction

Consider the classical problem of allocating an object to one of two agents whose valuations (i.e., willingness to pay in terms of a numéraire good) are *interdependent*, so that one agent's willingness to pay depends on the other agent's willingness to pay. The standard approach to modelling such interdependent preferences is to let each agent have a set of possible "types," where each type has a belief about the other agent's type, and each agent's valuation depends on both agents' types. A "direct mechanism" has each agent report his type, with the allocation of the object and agents' payments functions of the reported types. There are elegant characterizations of when there exists a direct mechanism where truth-telling is an equilibrium. But a usual criticism of this approach is that the type space is a construct of the modeler and it is not clear what it means to ask an agent to report a "type" that cannot be expressed in a natural language. Researchers have therefore devoted considerable effort to identifying "detail-free," indirect, mechanisms - where agents are asked to report something meaningful - that achieve the same objective.[1]

The objective of this paper is to provide a canonical description of possible interdependent preference types that is detail free in the sense that types' interdependent preferences can be described in a natural language without self-reference. We also want the definition to be operational in the sense that two distinct types are guaranteed to behave differently in at least one strategic setting. We consider a general setting with arbitrary finite sets of agents and outcomes. Agents are assumed to have expected utility preferences over lotteries. A mechanism specifies a finite set of possible messages for each agent and a lottery over outcomes for each message profile. We study equilibrium behavior. We identify a universal type space of possible interdependent (expected utility) preferences, expressed in a natural language, where distinct types in the universal type space are "strategically distinguishable" in the sense that there exist mechanisms where those types are guaranteed to behave differently in equilibrium. Our approach does not make a distinction between different reasons for the interdependence, such as whether it is based on objective signals, subjective assessments or psychological/behavioral reasons.[2]

We describe an agent's interdependent preference by a "first order preference" - in the example, giving the agent's willingness to pay for the object unconditional on the other agent's valuation; a "second order preference" - in the example, giving both (i) the agent's beliefs about the other agent's unconditional willingness to pay; and (ii) the agent's valuation of the object conditional on the other agent's first order preference; and so on... This gives a hierarchical description of the agent's

[1] Dasgupta and Maskin (2000) and Perry and Reny (2002) provide detail-free mechanisms for more general allocation problems.

[2] We discuss an example of psychological interdependence, and how it is incorporated, in the next Section.

interdependent preferences in a natural language. In the example, the set of all such hierarchies of interdependent preferences forms a universal type space of interdependent preferences, where each agent's type consists of (i) a belief over the other agent's type; and (ii) a mapping specifying the agent's valuation of the object as a function of the other agent's type.

But if we normalize valuations to be in the interval $[0,1]$, it is convenient, as a modelling device, to imagine that there are two private states for each agent corresponding to his "true value" of 0 or 1; and the agent faces uncertainty about his "true value." In this case, his set of possible valuations is isomorphic to the set of probability distributions over $\{0,1\}$. With this re-interpretation, our universal type space of interdependent preferences is formally equivalent to a version of the classic Mertens and Zamir (1985) universal type space of beliefs, but where beliefs are about private states (corresponding to extreme points of possible preferences) and in particular, an agent's beliefs about the other agent's private states are not relevant or recorded. Thus, our universal type space is a set T^* which is homeomorphic to the set of probability distributions over $\{0,1\} \times T^*$. An agent's belief over the other agent's type is then given by the marginal belief on the set T^*. The agent's valuation of the object conditional on the other agent's type is represented by the probability of state 1 conditional on the other agent's type. The subtlety in identifying the "right" universal type space of interdependent preferences consists of identifying what is the "right" state space over which the universal type space of beliefs is constructed: it consists of extreme points of a set of possible preferences. We illustrate this subtlety in the next Section by pointing out why alternative apparently intuitive approaches - including those usually explicitly or implicitly appealed to - fail to satisfy our criteria of providing a natural language and characterizing strategic distinguishability.

The strategic distinguishability question is due to Abreu and Matsushima (1992b) (AM). AM 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 (as well as equilibrium). 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. Lemma 2 of AM shows that two types are strategically distinguishable if and only if they differ in their preference hierarchies. Thus our main result is an extension of Lemma 2 in AM from finite to infinite type spaces (our result works for iterated deletion of strictly dominated strategies as well as equilibrium).[3] Our results make two main contributions relative to AM. First, our defin-

[3] 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 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. Bergemann and Morris (2009) and Brooks (2014) consider more special virtual implementation problems using ideas from Abreu and Matsushima (1992b) and this paper.

ition of universal types is independent of the type space in which they live. Second, AM require a distinct mechanism for each finite type space, even if we hold fixed the preference hierarchies we are trying to distinguish in those finite type spaces; by contrast, for any fixed pair of distinct preference hierarchies, we give a single finite mechanism that will strategically distinguish types with those preference hierarchies across all finite or infinite type spaces satisfying a continuity condition.

With the private states interpretation, our results show that we can strategically distinguish any pair of types if and only if they have distinct belief hierarchies about private states. Dekel, Fudenberg, and Morris (2006), (2007) (DFM) consider a setting where agents have beliefs and higher order beliefs about a common "external" state. They ask when and how we can distinguish types based on arbitrary games where agents' payoffs depend on the external state. It is an implication of their results that two types can be strategically distinguished if and only if they have distinct belief hierarchies about the external state.[4] Thus our results parallel those of DFM, where our uncertainty is about agents' (perhaps interdependent) payoffs from fixed outcomes, rather than about the external state that will determine payoffs. We present an extension of our results, where we add external states that mechanisms can be made contingent on. This generalization not only embeds the DFM results in the original form, but also implies that the DFM results would go through if we restricted attention to special classes of games such as zero sum games or common interest games.[5]

Thus our results can be seen as a unification and extension of the results of AM and DFM. Like DFM and unlike AM, we express types in a detail-free way and distinguish types on arbitrary type spaces. Like AM and unlike DFM, we distinguish between types with different hierarchies of preferences and not just types with different hierarchies of beliefs about external states, and thus we are more constrained in the set of strategic settings we can confront agents with.

The extension requires an innovation in the proof strategy. In order to strategically distinguish two types with distinct preference hierarchies, we construct a mechanism in which agents are asked to report their preference hierarchy. For each agent i and each order n, there is a positive probability that a component of the mechanism is selected where outcomes are chosen to give agent i an incentive to truthfully report her nth order preference conditional on other agents' having truthfully reported their $(n-1)$th and lower order preferences. A potential difficulty with this proof strategy is that agent i's report of her nth order preference is an input not only into the component giving her an incentive to truthfully report her nth order preference, but also into the components giving the other agents incentives to truthfully report their $(n+1)$th and higher order preferences. Abreu and Matsushima (1992b) dealt with this difficulty by exploiting finiteness, and

[4] We discuss exactly which DFM results imply this in Section 4.5.

[5] Gossner and Mertens (2001) suggested such a result for zero sum games.

having the probability of all components about $(n+1)$th and higher order preferences occur with much smaller probability than components involving nth order preferences. Dekel, Fudenberg, and Morris (2006) can choose payoffs so that components giving each agent an incentive to report her preferences have no implications for other agents. Neither trick is available in our setting, as we have arbitrary type spaces and agents' preferences over outcomes may be arbitrarily linked. Instead, we develop a *robust scoring rule* that not only gives an agent an incentive to report her nth order preferences truthfully if others report their $(n-1)$th and lower order preferences truthfully, but also gives the agent an incentive to report her nth order preferences *approximately* truthfully if others report their $(n-1)$th and lower order preferences approximately truthfully. This enables us to design a mechanism where the error size at each of a finite number of orders can be simultaneously controlled.[6]

Our results hold only if we exclude agents who are completely indifferent over all outcomes. Clearly, completely indifferent types cannot be strategically distinguished from any other types. We must also impose a continuity restriction on agents' preferences. Thus in the single good allocation example, it is necessary that the agents' valuations are in a bounded interval. Without such restrictions, we show that, while it is possible to find out agents' first order preferences (in the example, their unconditional valuations of the object), it is not possible to strategically distinguish two types that have the same first order preference. In the main body of the paper, we impose continuity by assuming "simplex" restrictions on preferences, so that an agent's ex post preference over outcomes, conditional on any event that may occur, can be uniquely represented as a convex combination of a finite set of possible preferences over lotteries. The simplex assumption allows us to develop our results in as standard language as possible, to draw tight and transparent connections with the universal type space of Mertens and Zamir (1985), and to present our work as a generalization and unification of Abreu and Matsushima (1992b) and Dekel, Fudenberg, and Morris (2006), (2007). Generalizing beyond the simplex case requires a more direct language for preference hierarchies, as they cannot be simply expressed as belief hierarchies over private states. Using this less standard but more direct alternative language (which is of independent interest), we also show that our main equilibrium strategic distinguishability result generalizes beyond the simplex case to weaker ("λ-continuity") restrictions requiring that agents' preferences are uniformly bounded away from complete indifference.[7]

Our focus in this paper is on when two types are "strategically indistinguishable," so that, in

[6] A related issue arises in the work of Chambers and Lambert (2014), where the problem of eliciting dynamic (rather than interactive) beliefs is studied.

[7] This generalization is also needed to present Lemma 2 of Abreu and Matsushima (1992b) as a formal special case of our results.

any mechanism, there is a common equilibrium action which each of them might take. A more demanding requirement is that two types are "strategically equivalent," so that, in any mechanism, the set of actions they might play are the same. Exploiting the connection with DFM, another result that we can show is that two types are strategically equivalent under the solution concept of interim correlated rationalizability (ICR) if and only if they have the same preference hierarchy (or belief hierarchy over private states). We can then use this result to establish strategic distinguishability for interim correlated rationalizability, for equilibrium and for any solution concept which is finer than ICR and coarser than equilibrium. However, this strategic equivalence result (unlike our equilibrium and rationalizability strategic distinguishability results) depends on the simplex assumption and does not extend to more general types spaces; this is an implication of results reported in Morris and Takahashi (2012) and is not explored in this paper.

The paper is organized as follows. In Section 2, we return to the example where agents' ex post preferences are parameterized by a single number in the interval $[0, 1]$, and motivate why the universal type space that we construct is the right one by spelling out why intuitive alternatives are either too large or too small in terms of our objectives of providing a natural language and characterizing strategic distinguishability. In Section 3, we present our main equilibrium strategic distinguishability result when a simplex restriction is imposed on ex post preferences, using a strategic equivalence result for rationalizability to prove the result. In Section 4, we discuss further connections to the literature, including what happens when non-expected utility preferences are permitted (Epstein and Wang (1996)), to what extent we could have separated the treatment of belief types and ex post preference types (Gul and Pesendorfer (2010)), and the relation to classical revealed preference theory (Afriat (1967)); we also discuss how allowing uncertainty about external observable states enables us to make an exact connection with the work of Dekel, Fudenberg, and Morris (2006), (2007). In Section 5, we report why continuity restrictions are necessary for our results and report the weakest continuity restriction we know under which our main results go through.

2 The "Right" Universal Type Space

Before introducing our formal framework, we give a motivating example to illustrate why we formalize the problem the way we do, and give intuition for our results. We first describe three very different scenarios where each agent's preference is summarized by a single payoff parameter in the interval $[0, 1]$. We then motivate the mathematical description of the universal type space which we propose independent of the interpretation of the payoff parameter. Finally, we explain why alternative universal type spaces implicitly or explicitly proposed in the literature do not serve our

purposes.

Thus suppose there are two agents, and each agent i has a payoff parameter r_i in the interval $[0, 1]$. This can be used to represent the scenario in the introduction: an agent's payoff parameter represents his willingness to pay for an object in terms of a numéraire good. Call this the "private good" scenario.

For an alternative scenario, consider "conditional altruism."[8] Each of two agents may care about the other's private consumption, so that he is *altruistic*. But an agent may also be more altruistic if he thinks that the other agent is more altruistic, in which case he is *conditionally altruistic*. Higher order conditional altruism is also a possibility. In this case, agents' preferences are interdependent for *psychological* reasons. More concretely, suppose that a prize is being allocated to either of the two agents. Suppose that there is a probability $r_i \in [0, 1]$ such that agent i is indifferent between the other agent getting the object for sure and getting the object himself with probability r_i. Thus r_i is an index of the agent i's altruism. Conditional altruism corresponds to having a higher altruism index when the other agent has a higher index.

Finally, consider a third "cardinality" scenario. There are three outcomes, "best," "medium" and "worst." There is common certainty of agents' common ordinal preferences. Thus both agents strictly prefer "best" to "worst" and both agents weakly prefer "best" to "medium" and "medium" to "worst." Their expected utilities preferences over lotteries on the three outcomes may differ, however. An agent with preferences parameterized by $r_i \in [0, 1]$ is indifferent between outcome "medium" for sure and a lottery with probability r_i on "best" and $1 - r_i$ on "worst." Thus if we normalize the von Neumann-Morgenstern utility index of "best" to 1 and of "worst" to 0, then r_i measures the utility index of outcome "medium."

We will propose the same universal type space of interdependent preferences, modulo the interpretation of the payoff parameters r_1 and r_2, in the three scenarios.[9] Agent i's interdependent preference type will have the following hierarchical description:

1. A first order preference given by payoff parameter r_i describing the agent's unconditional preference.

2. A second order preference over Anscombe-Aumann acts giving outcomes as functions of the other agent's first order preference.

3. A third order preference over Anscombe-Aumann acts giving outcomes as functions of the other agent's second order preference.

[8] This discussion follows Levine (1998) and Gul and Pesendorfer (2010).

[9] We postpone until Section 3.2 a discussion of how these scenarios will fit formally into our framework.

4. And so on....

In the private good scenario, a second order preference will give both (i) the agent's belief about the other agent's unconditional valuation and (ii) specify the agent's valuation conditional on possible unconditional valuations of the other agent. In the conditional altruism scenario, a second order preference consists of (i) a belief about the unconditional altruism of the other agent, as well as (ii) specifying the agent's altruism conditional on the other agent's unconditional altruism. In the cardinality scenario, a second order preference consists of (i) a belief about the other agent's cardinal valuation of the medium outcome, as well as (ii) specifying the agent's valuation of the medium outcome conditional on the other agent's cardinal valuation.

We claim that - independent of the scenarios - the "right" universal type space for each agent in this symmetric environment is the universal type space T^* of beliefs based on $\{0,1\}$, i.e., the set of coherent belief hierarchies about the extreme points of own payoff parameters, associated with the following homeomorphism

$$T^* \cong \Delta(\{0,1\} \times T^*).$$

Thus each agent's type is uniquely identified with a belief over $\{0,1\} \times T^*$. The interpretation is that the marginal belief on T^* corresponds to the agent's belief over the other agent's type; and the conditional probability of state 1 corresponds to the agent's payoff parameter conditional on the other agent's type.

A useful way to see why it is the right space for our purposes is to consider three natural alternative spaces sometimes implicitly or explicitly proposed and explain why they do not succeed in our purposes of (i) expressing interdependent preference types in a natural language; and (ii) characterizing strategic distinguishability.

First, we could identify types with their own payoff parameters and their beliefs over other agents' types. For reasons that we will explain, we will call this the "*private values* (PV) *universal type space*." The private values universal type space T_{PV} will satisfy the homeomorphism

$$T_{PV} \cong [0,1] \times \Delta(T_{PV}).$$

The space T_{PV} is not large enough to express the interdependent preferences we are interested in. It clearly identifies a first order preference. But it does not allow second order preferences to depend on the other agent's first order preference. In other words, it rules out the interdependence we are trying to capture and is thus too small. This space would be the "right" universal type space if common certainty of private values were maintained.[10] It corresponds to a strict subset of our universal type space T^* where beliefs about $\{0,1\}$ and T^* are constrained to be independent.

[10] See, for example, Heifetz and Neeman (2006) for a general construction of such a private value universal type space.

Second, we could identify types with coherent belief hierarchies about all agents' payoff parameters. Call this the "*payoff* (P) *universal type space*." In this example, agents' ex post preferences are parameterized by a profile $(r_1, r_2) \in [0,1]^2$. The payoff universal type space will satisfy the homeomorphism

$$T_P \cong \Delta\left([0,1]^2 \times T_P)\right).$$

The space T_P is large enough to express all interdependent preferences that we are interested in, but it is too rich for our purposes. The language distinguishes between agent i being sure that his payoff parameter is $\frac{1}{2}$ and having a 50/50 belief about where his parameter is 0 or 1. But this distinction is not meaningful in a natural language description of the agent's preferences: in the private good scenario, in each case, the agent has a willingness to pay for the object of $\frac{1}{2}$. If we did allow such a richer language where these situations were treated differently, it would not be possible to strategically distinguish such types. Similarly, here an agent has a belief not only about his own payoff parameter, but also about the other agent's payoff parameter. Thus in the conditional altruism scenario, we allow for the possibility that agent i is sure that agent j is "truly altruistic" (i.e., $r_j = 1$) even though i is sure that j is sure that j is "truly selfish" (i.e., $r_j = 0$), and j will never behave in an altruistic way. Again, it is not clear what i's belief about j's payoff parameter means, and if we allowed such a rich language, such a type would not be strategically distinguishable from agent i who is sure that $r_j = 0$.

Third, we could identify types with beliefs and higher order beliefs about a rich class of "payoff types" that described interdependence in ex post preferences. An agent knows his own payoff type but may not know the other agent's payoff type. We will call this the "*interdependent payoff* (IP) *universal type space*." Thus we may let Θ be a set of possible payoff types for each agent and let $r\left(\theta, \theta'\right) \in [0,1]$ specify an agent's payoff parameter when he has payoff type θ and the other agent has payoff type θ', where $r : \Theta^2 \to [0,1]$. The interdependent payoff universal type space T_{IP} will satisfy the homeomorphism

$$T_{IP} \cong \Theta \times \Delta(T_{IP}).$$

Since we assumed that agents knew their own "payoff types," this is simply the private values universal type space defined over Θ instead of $[0,1]$ as we did for T_{PV}. This modelling approach follows a standard practise in the literature of treating payoff interdependence and higher order beliefs separately, and is widely used in the mechanism design literature, either implicitly or explicitly. It is implicit in Dasgupta and Maskin (2000), who introduce "types" which determine players' interdependent values and then consider ways of implementing the efficient outcome that do not depend on beliefs. It is explicit in the work of two of us on robust mechanism design (Bergemann and Morris (2012)), where we assumed a space of possible "payoff types," and allow any beliefs and

higher order beliefs about those payoff types.

The payoff type spaces in Dasgupta and Maskin (2000) and Bergemann and Morris (2012) are not intended to be universal. Gul and Pesendorfer (2010) constructed a universal type space of interdependent ex post preferences, abstracting from any belief structure. In particular, they identify a maximal set of interdependent payoff types which captures all distinctions that can be expressed in a natural language. When they consider applications of their universal type space to incomplete information settings, they treat incomplete information separately and thus implicitly allow any beliefs and higher order beliefs over their universal payoff space.

Similarly to the space T_P, the space T_{IP} is large enough to express interdependent preferences if the underlying payoff type space is rich enough, but it is then too rich for our purposes. An agent's type in T_{IP} specifies what his beliefs would be conditional on types of the other agents he attaches probability zero to. Thus it contains counterfactual information. While there might be purposes for which we want a language to express this information, as discussed in Gul and Pesendorfer (2010), such distinctions will not be strategically distinguishable in our sense.[11]

Concretely, in the conditional altruism scenario, compare (i) an agent who shares with the other agent because he is conditionally altruistic and is sure that the other agent is altruistic; and (ii) an agent who shares with the other agent because he is unconditionally altruistic. In the private good scenario, compare (i) an agent with interdependent values who is certain that a good is worth $\frac{1}{2}$ because he is sure that the other agent has observed a good signal, and (ii) an agent with a private value of $\frac{1}{2}$ for the good. These agents will not be strategically distinguishable from each other, but will correspond to different types in T_{IP}.

3 Main Result

We let I be a non-empty finite set of agents, and Z be a finite set of outcomes with $|Z| \geq 2$. We will impose "simplex" restrictions on the possible ex post preferences of each agent. The interpretation is that there is common certainty that each agent's preference is within that set, conditional on any event. To keep language as standard as possible, we find it convenient to identify expected utility preferences with representations of those preferences in $\mathbb{R}^Z$. Thus an agent with preference $u_i \in \mathbb{R}^Z$ prefers lottery $p \in \Delta(Z)$ to lottery $p' \in \Delta(Z)$ if

$$\sum_{z \in Z} p(z) u_i(z) \geq \sum_{z \in Z} p'(z) u_i(z).$$

[11] Our notion of strategic distinguishibility concerns static games and static solution concepts. In dynamic games, with sequentially rational solution concepts, counterfactuals are relevant for strategic analysis. In this case, a richer type space, perhaps following the work of Battigalli and Siniscalchi (1999), would be required.

3.1 Simplex Restrictions

For each agent $i \in I$, we consider a convex polytope $\overline{U}_i \subset \mathbb{R}^Z$ with vertices (i.e., extreme points) $U_i = \{u_i^1, u_i^2, \ldots, u_i^{K_i}\}$ such that

1. no two distinct utility indices in $\overline{U}_i$ represent the same preference, i.e., if there exist $u_i, v_i \in \overline{U}_i$, $\alpha > 0$, and $\beta \in \mathbb{R}$ such that $v_i(z) = \alpha u_i(z) + \beta$ for all $z \in Z$, then $u_i = v_i$;

2. no utility index in $\overline{U}_i$ represents the complete indifference, i.e., $u_i \neq (x, \ldots, x)$ for any $u_i \in \overline{U}_i$ and any $x \in \mathbb{R}$;

3. $u_i^2 - u_i^1, u_i^3 - u_i^1, \ldots, u_i^{K_i} - u_i^1$ are linearly independent.

Property (1) is simply a normalization among representations. Property (2) rules out the possibility of complete indifference within the set. Property (3) is the simplex assumption; it requires that every preference in $\overline{U}_i$ can be uniquely represented as a convex combination of the extreme points. We will discuss further below how this assumption may fail and what the implications of its failure are. We will hold a profile of simplex restrictions $\left(\overline{U}_i\right)_{i \in I}$ fixed throughout our analysis in the next two Sections.

Our interpretation is that $\overline{U}_i$ represents a set of preferences. Each $\overline{U}_i$ is isomorphic to the set of probability distributions over its extreme points, $\Delta(U_i)$, and this will play an important role in our presentation. A preference can then be thought of as a probability distribution over "private states" U_i. We find it useful to sometimes use this language and interpretation (identifying ex post preferences with a probability distribution over private states) in the analysis, although we emphasize that these subjective states have no observable counterpart.

3.2 Illustrating Simplex Restrictions

Simplex restrictions can be illustrated by noting how the examples discussed in Section 2 fit into the framework of this Section. Suppose that there are two agents. First, consider the cardinality scenario. In this case, there were three outcomes, which we can write as $Z = \{w, m, b\}$, corresponding respectively to "worst," "medium" and "best" outcomes. The set U_i is the same for each agent i, consisting of the two vectors $(0, 0, 1)$ and $(0, 1, 1)$ corresponding to the extreme preferences where the agent is indifferent between the medium outcome m and the worst outcome w and where the agent is indifferent between the medium outcome m and the best outcome b.

For the conditional altruism scenario, we can again consider three outcomes, $Z = \{\varnothing, 1, 2\}$, but where the outcomes now correspond to, respectively, no one getting the prize, agent 1 getting the prize and agent 2 getting the prize. Now U_1 is different from U_2, but has the same structure.

The set U_1 consists of two vectors $(0,1,0)$ and $(0,1,1)$ corresponding to, respectively, the extreme preferences where the agent 1 is indifferent between the other agent getting the prize (outcome 2) and no one getting the prize (outcome $\varnothing$) and where the agent is indifferent between the other agent getting the prize (outcome 2) and getting the prize himself (outcome 1). Symmetrically, U_2 consists of two vectors $(0,0,1)$ and $(0,1,1)$.

To fit the private good scenario into our finite outcome setting, we must introduce outcomes that play the role of a numéraire. Suppose that each agent can be allocated nothing ($\varnothing$), a known good (k) or a new good of uncertain value (u). Each agent cares only about his private allocation. Each agent prefers the known good to the unknown good, but his valuation of the unknown good may be anywhere between getting nothing and getting the known good. Thus his valuation of the unknown good corresponds to his willingness to pay for the new good in terms of probability of getting the known good. Formally, the set of outcomes specifies one of the three private outcomes for each agent, so

$$Z = \{\varnothing, k, u\}^2 = \{\varnothing\varnothing, \varnothing k, \varnothing u, k\varnothing, kk, ku, u\varnothing, uk, uu\},$$

where, for example, $u\varnothing$ corresponds to agent 1 getting the new good and agent 2 getting nothing. The set U_1 consists of two vectors:

$$(0,0,0,1,1,1,0,0,0)$$

and

$$(0,0,0,1,1,1,1,1,1),$$

corresponding to the extreme preferences where the agent 1 is indifferent between the unknown good (outcomes with $z_1 = u$) and getting nothing (outcomes with $z_1 = \varnothing$) and where the agent is indifferent between the unknown good and getting the known good (outcomes with $z_1 = k$). Symmetrically, U_2 consists of two vectors:

$$(0,1,0,0,1,0,0,1,0)$$

and

$$(0,1,1,0,1,1,0,1,1).$$

We can also use these examples to illustrate how restrictive the simplex assumption is. For example, we can enrich these examples to allow more uncertainty about the ranking of outcomes. More specifically, in the conditional altruism scenario, we could replace the two extreme preferences of agent 1 by $(0,1,v)$ and $(0,1,-v)$, for some large $v > 0$. This allows for the possibility that agent 1 strictly prefers agent 2 getting the prize to getting the prize himself. And it allows a "spiteful"

agent 1 who strictly prefers no one getting the prize to agent 2 getting the prize. This continues to be a simplex case.

But now suppose instead that we added a fourth outcome m' to the cardinality scenario, so $Z = \{w, m, m', b\}$. Let outcome m' also be "medium" in the sense that outcome b is always weakly preferred to m' and m' is weakly preferred to w. Also suppose that we have no restriction on how m is ranked relative to m'. These preferences are naturally represented by the convex hull of four vectors $(0,0,0,1)$, $(0,0,1,1)$, $(0,1,0,1)$ and $(0,1,1,1)$. This set of preferences is not a simplex. For example, the preference $\left(0, \frac{1}{2}, \frac{1}{2}, 1\right)$ does not have a unique representation as a convex combination of extreme points. It is a 50/50 combination of $(0,0,0,1)$ and $(0,1,1,1)$. But it is also a 50/50 combination of $(0,0,1,1)$ and $(0,1,0,1)$.

3.3 Type Spaces

A *type space* $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ consists of non-empty measurable spaces T_i of agent i's possible types, $T_{-i} \equiv \prod_{j \neq i} T_j$, and measurable mappings:

$$\mu_i \colon T_i \to \Delta(U_i \times T_{-i})$$

that assign each type $t_i \in T_i$ with a belief $\mu_i(\cdot \mid t_i)$. Now for any Anscombe-Aumann acts $f, f' : T_{-i} \to \Delta(Z)$, type t_i prefers f to f' if

$$\int_{U_i \times T_{-i}} \sum_{z \in Z} u_i(z) \left(f(z \mid t_{-i}) - f'(z \mid t_{-i}) \right) \mu_i(du_i, dt_{-i} \mid t_i) \geq 0.$$

We interpret the marginal probability distribution $\operatorname{mrg}_{T_{-i}} \mu_i(\cdot \mid t_i) \in \Delta(T_{-i})$ as type t_i's belief over the opponents' type profiles, and the conditional probability distribution $\mu_i(\cdot \mid t_i, t_{-i}) \in \Delta(U_i) \cong \overline{U}_i$ as the utility index that represents type t_i's ex-post preference when the opponents' types are t_{-i}. Note that correlation in $\Delta(U_i \times T_{-i})$ is essential, as it allows us to express interdependency of type t_i's preferences and the opponents' types t_{-i}.

3.4 The Universal Type Space

Because ex post preferences $\overline{U}_i$ are isomorphic to $\Delta(U_i)$, we can also interpret U_i as a finite set of extreme "payoff states" and $u_i \in \overline{U}_i$ as a probability distribution over payoff states. Thus we can treat $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ formally as a belief type space, where different agents have beliefs and higher-order beliefs over private state spaces $(U_i)_{i \in I}$. Thus, with minor modifications of Mertens and Zamir (1985) and Brandenburger and Dekel (1993), we define T_i^* as the set of all coherent hierarchies of agent i's beliefs, and construct the universal type space $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i \in I}$ with homeomorphism

$$\mu_i^* \colon T_i^* \to \Delta(U_i \times T_{-i}^*).$$

Moreover, for any type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$, the mapping that maps each type in T_i to its hierarchy of beliefs:

$$\hat{\mu}_i \colon T_i \to T_i^*$$

preserves the belief structure, i.e.,

$$\mu_i^*(E \mid \hat{\mu}_i(t_i)) = \mu_i(\{(u_i, t_{-i}) \mid (u_i, \hat{\mu}_{-i}(t_{-i})) \in E\} \mid t_i)$$

for any measurable subset $E \subseteq U_i \times T_{-i}^*$. We sometimes write $\hat{\mu}_i(\cdot; \mathcal{T})$ to emphasize its domain.

3.5 Rationalizability, Strategic Equivalence and Strategic Distinguishability

A (finite) *mechanism* (or *game form*) is given by $\mathcal{M} = ((M_i)_{i \in I}, O)$, where M_i is a non-empty finite set of messages (actions) available to agent i, and $O \colon M \equiv \prod_i M_i \to \Delta(Z)$ is the outcome function. In this mechanism, agents send messages $m \in M$ simultaneously, and the mechanism assigns an outcome z with probability $O(z \mid m)$. As usual, the domain of O is extended to $\Delta(M)$ in the multi-linear way.

A type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ and a mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$ together define a game of incomplete information $(\mathcal{T}, \mathcal{M})$. We will later define and discuss equilibrium and other solution concepts for this game of incomplete information. However, it is useful to first discuss a suitably modified definition of solution concept of interim correlated rationalizability in Dekel, Fudenberg, and Morris (2007) to this setting. The difference from those papers is one of structure - here there are private states $(U_i)_{i \in I}$ instead of common states - as well as interpretation; here, the states represent extreme points of the agents' possible ex post preferences, while in DFM, they represented external events on which games' payoffs were conditioned.[12] We define rationalizability as follows: the induction is initialized with:

$$R_i^0(t_i) = M_i,$$

the inductive step $n+1$ is defined by:

$$R_i^{n+1}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists a measurable mapping } \sigma_{-i} \colon U_i \times T_{-i} \to \Delta(M_{-i}) \text{ s.t.:} \\ \text{(i) } \sigma_{-i}(R_{-i}^n(t_{-i}) \mid u_i, t_{-i}) = 1 \text{ for any } u_i \in U_i \text{ and } t_{-i} \in T_{-i}, \\ \text{(ii) } \int_{U_i \times T_{-i}} \sum_{m_{-i} \in M_{-i}} \sum_{z \in Z} u_i(z)(O(z \mid m_i, m_{-i}) - O(z \mid m_i', m_{-i})) \\ \qquad \times \sigma_{-i}(m_{-i} \mid u_i, t_{-i}) \mu_i(du_i, dt_{-i} \mid t_i) \geq 0, \text{ for any } m_i' \in M_i. \end{array} \right\},$$

[12] Allowing correlation in an agent's conjecture about that agent's private state and another agent's action allows correlation of preferences with others' actions. In this sense, ICR can be seen as even more permissive in this context. See Morris and Takahashi (2012) for more on the foundations and interpretations of such solution concepts.

which is equivalent to

$$R_i^{n+1}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \nu_i \in \Delta(M_{-i} \times U_i \times T_{-i}) \text{ s.t.:} \\ \text{(i) } \nu_i(\{(m_{-i}, u_i, t_{-i}) \mid m_{-i} \in R_{-i}^n(t_{-i})\}) = 1, \\ \text{(ii) } \operatorname{mrg}_{U_i \times T_{-i}} \nu_i = \mu_i(t_i), \\ \text{(iii) } \int_{M_{-i} \times U_i \times T_{-i}} \sum_{z \in Z} u_i(z)(O(z \mid m_i, m_{-i}) - O(z \mid m_i', m_{-i})) \\ \qquad \times \nu_i(dm_{-i}, du_i, dt_{-i}) \geq 0, \text{ for any } m_i' \in M_i. \end{array} \right\},$$

and the limit set is defined by:

$$R_i(t_i) = \bigcap_{n=0}^{\infty} R_i^n(t_i).$$

We sometimes write $R_i(t_i; \mathcal{T}, \mathcal{M})$ to emphasize the underlying type space $\mathcal{T}$ and mechanism $\mathcal{M}$. Similarly to Dekel, Fudenberg, and Morris (2007, Proposition 1 and Corollary 2), we can show that rationalizability depends only on universal types.

Proposition 1 *For any type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$, any agent $i \in I$, and any type $t_i \in T_i$, we have*

$$R_i(t_i; \mathcal{T}, \mathcal{M}) = R_i(\hat{\mu}_i(t_i); \mathcal{T}^*, \mathcal{M})$$

for any mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$.

The proof of this Proposition appears in the Appendix.

Following the terminology in the introduction, for the ICR solution concept, we say that two types are *strategically indistinguishable* if their ICR actions have a non-empty intersection for every mechanism $\mathcal{M}$. We say that the types are *strategically equivalent* if there exists a mechanism $\mathcal{M}$ such that their ICR action sets are the same. In this terminology, the following Theorem establishes that the universal type space characterizes both strategic distinguishability and strategic equivalence for ICR.

Theorem 1 *For any two type spaces $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ and $\mathcal{T}' = (T_i', \mu_i')_{i \in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t_i' \in T_i'$, the following three conditions are equivalent:*

1. *$\hat{\mu}_i(t_i; \mathcal{T}) = \hat{\mu}_i(t_i'; \mathcal{T}')$;*
2. *$R_i(t_i; \mathcal{T}, \mathcal{M}) \cap R_i(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$ for any mechanism $\mathcal{M}$;*
3. *$R_i(t_i; \mathcal{T}, \mathcal{M}) = R_i(t_i'; \mathcal{T}', \mathcal{M})$ for any mechanism $\mathcal{M}$.*

Note that $1 \Rightarrow 3$ follows from Proposition 1; $3 \Rightarrow 2$ follows from the nonemptyness of rationalizability. Then $2 \Rightarrow 1$ follows from the following Proposition. Let d_i^* be a metric compatible with the product topology on T_i^*.

Proposition 2 *For every $\varepsilon > 0$, there exists a mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$ such that*

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

for any two type spaces $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ and $\mathcal{T}' = (T_i', \mu_i')_{i \in I}$ based on $(U_i)_{i \in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t_i' \in T_i'$.

The Proposition proves a little more than what is needed to prove Theorem 1: it shows that if we fix a metric d_i^* and $\varepsilon > 0$, we can strategically distinguish all types that are at least ε apart using the *same mechanism*. In the remainder of this sub-Section, we describe the mechanism we use to prove this result, which is the main technical contribution of the paper. Proofs are in the Appendix.

The strategy of proof is as follows. If two types are ε apart in the metric compatible with the product topology on T_i^*, then there must exist $\overline{\varepsilon} > 0$ and N such that the types Nth order preferences are at least $\overline{\varepsilon}$ apart. For each order $n = 0, 1..., N$, we will choose an accuracy $\varepsilon_n > 0$. For each agent i and $n \geq 1$, agent i will report an element of a ε_{n-1}-dense finite subset of his nth order preference. For each agent i and $n \geq 1$, there will be a component of the mechanism, chosen with positive probability, that will pick an outcome as a function of agent i's announcement of his nth order preference and the announcements of the $(n-1)$th and lower order preferences of the other agents. The mechanism will have the property that as long as $(n-1)$th and lower order announcements are within ε_{n-1} of the truth, then nth order announcements are within ε_n of the truth.

The last step of the argument uses a robust scoring rule described the next sub-Section. We show that, for every $\varepsilon > 0$, we can find $\delta > 0$ and a scoring rule that gives an incentive to report beliefs within ε of his true beliefs even if the outcomes of the scoring rule may be arbitrarily perturbed within δ. This lemma can then be iteratively applied to construct the mechanism used in the main proof.

Abreu and Matsushima (1992b) and Dekel, Fudenberg, and Morris (2006) followed similar arguments up until the last step. But neither required the robust scoring lemma. AM exploited the finiteness of the type space. They can choose an $\varepsilon > 0$ such that the $(j, n+1)$th component occurs with probability at most ε times that of the (i, n)th component. Now $\varepsilon > 0$ can be chosen uniformly small enough so that agents can be strictly incentivized to report their preferences exactly at every order.[13] DFM allow for arbitrary, infinite, type spaces, so it is not possible to find a uniform ε that makes AM argument work. In DFM, it is necessary to have agents report an approximation

[13] In the related work of Bergemann and Morris (2009), there is a finite set of possible "payoff types" and an analogous trick can be applied.

to their true belief at every order. But payoffs for each agent can be chosen independently, so it is possible to do the approximation one order at a time. Because neither strategy is available in our setting, we need a novel robust scoring rule to make the argument work.

3.5.1 The Robust Scoring Rule and the Proof of Proposition 2

As a preliminary step, we first analyze a single-agent mechanism that reveals her state-dependent preferences. In this subsection, fix a simplex $\overline{U} \subset \mathbb{R}^Z$ of ex post preferences with vertices U and a compact metric space X of states with metric d. Let d_Δ be a metric compatible with the weak-* topology over $\Delta(U \times X)$. Let $F(X)$ be the set of (Anscombe-Aumann) acts over X, i.e., the set of measurable functions $f\colon X \to \Delta(Z)$. Then each $\mu \in \Delta(U \times X)$ uniquely represents a state-dependent preference over $F(X)$. That is, the agent with preference μ weakly prefers f to f' if and only if

$$\int_{U\times X} \sum_{z\in Z} u(z)(f(z \mid x) - f'(z \mid x))\mu(du, dx) \geq 0.$$

We define the choice function with respect to μ:

$$C_\mu(f, f') = \begin{cases} f & \text{if } \mu \text{ weakly prefers } f \text{ to } f', \\ f' & \text{if } \mu \text{ strictly prefers } f' \text{ to } f, \end{cases}$$

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 $\mathcal{M}^0 = (M^0, O^0)$ for a single agent with message set $M^0 = \Delta(U \times X)$ and outcome function $O^0\colon M^0 \times X \to \Delta(Z)$ given by

$$O^0(z \mid m, x) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} 2^{-k-l} C_m(f_k, f_l)(z \mid x), \tag{1}$$

for each realized state $x \in X$ and reported preference $m \in M^0$. 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.[14]

In Lemma 1 below, we show that truth telling is optimal $\mathcal{M}^0$ for every type. Indeed, by invoking the compactness of X, we show a "robust" version of optimality: in every mechanism close to $\mathcal{M}^0$,

[14] Note that $\mathcal{M}^0$ is not a finite mechanism. The mechanism we will construct in the next Subsection to prove Proposition 2, however, has finite actions.

the agent strictly prefers reporting almost true preferences to reporting others according to almost true preferences.

Recall that for each message m, $O^0(\cdot \mid m, \cdot)$ is an act over X, which determines an outcome z with probability $O^0(z \mid m, x)$ 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 $O^0(\cdot \mid m, x)$. Formally, for each $\delta > 0$ and measurable space Ω, we consider perturbed outcome function $O\colon M^0 \times X \times \Omega \to \Delta(Z)$ such that $\|O(\cdot \mid \cdot, \cdot, \omega) - O^0\| \equiv \sup_{m \in M^0, x \in X, z \in Z} |O(z \mid m, x, \omega) - O^0(z \mid m, x)| \leq \delta$ for any $\omega \in \Omega$. Second, when nature is supposed to choose x, nature may instead choose x' in a neighborhood of x. Formally, for each $\delta > 0$, $\mu \in \Delta(U \times X)$, and measurable space Ω, let

$$\Delta_{\delta,\mu}(U \times X \times \Omega) = \left\{ \mathrm{mrg}_{1,3,4}\mu' \in \Delta(U \times X \times \Omega) \;\middle|\; \begin{array}{l} \exists \mu' \in \Delta(U \times X \times X \times \Omega) \text{ s.t.} \\ \text{(i) } \mu'(\{(u, x, x', \omega) \mid d(x, x') \leq \delta\}) = 1, \\ \text{(ii) } \mathrm{mrg}_{1,2}\mu' = \mu \end{array} \right\}, \quad (2)$$

where $\mathrm{mrg}_\Lambda \mu'$ with $\Lambda \subset \{1, 2, 3, 4\}$ denotes the marginal of μ' with respect to the coordinates in Λ. In words, $\Delta_{\delta,\mu}(U \times X \times \Omega)$ is the set of preferences over noisy acts induced by the original preference μ.

Lemma 1 *For every $\varepsilon > 0$, there exists $\delta > 0$ such that the following is true for any preference $\mu \in \Delta(U \times X)$, any pair of messages m, m', any measurable space Ω, and any perturbed outcome function $O\colon M^0 \times X \times \Omega \to \Delta(Z)$: if $d_\Delta(\mu, m) \leq \delta$, $d_\Delta(\mu, m') > \varepsilon$, and $\|O(\cdot \mid \cdot, \cdot, \omega) - O^0\| \leq \delta$ for any $\omega \in \Omega$, then any preference in $\Delta_{\delta,\mu}(U \times X \times \Omega)$ strictly prefers $O(\cdot \mid m, \cdot, \cdot)$ to $O(\cdot \mid m', \cdot, \cdot)$.*

Recall that we follow the standard procedure and construct the universal type space $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i \in I}$ by hierarchies. Specifically, for each $i \in I$, let $H_{i,0} = \{*\}$ be initialized with a single element, and let $H_{i,n} = H_{i,n-1} \times \Delta(U_i \times H_{-i,n-1})$ for each $n \geq 1$. Note that $H_{i,n} = \prod_{k=0}^{n-1} \Delta(U_i \times H_{-i,k})$. Then we can construct the universal type space $T_i^* \subset \prod_{n=0}^{\infty} \Delta(U_i \times H_{-i,n})$ as the set of coherent hierarchies of agent i's beliefs over his private states. Let $d_{i,n}$ be a metric compatible with the topology on the set of agent i's n-th order beliefs, $\Delta(U_i \times H_{-i,n-1})$.

Fix any $\varepsilon > 0$. By the definition of the product topology, there exist $\bar{\varepsilon} > 0$ and $N \in \mathbb{N}$ such that, for every $\{t_{i,n}\}_{n=1}^{\infty}, \{t'_{i,n}\}_{n=1}^{\infty} \in T_i^*$, if $d_i^*(\{t_{i,n}\}_{n=1}^{\infty}, \{t'_{i,n}\}_{n=1}^{\infty}) > \varepsilon$, then there exists some $n \leq N$ such that $d_{i,n}(t_{i,n}, t'_{i,n}) > \bar{\varepsilon}$. Pick such $\bar{\varepsilon}$ and N.

For each $i \in \mathcal{I}$ and $n \leq N$, we apply Lemma 1 by substituting $X = H_{-i,n-1} = \prod_{j \neq i} \prod_{k=0}^{n-2} \Delta(U_j \times H_{-j,k})$, $d(t_{-i,n-1}, t'_{-i,n-1}) = \max_{j \neq i} \max_{1 \leq k \leq n-1} d_{j,k}(t_{j,k}, t'_{j,k})$, and $d_\Delta = d_{i,n}$. Pick a countable dense subset of $F_c(H_{-i,n-1})$, and define $O^0_{i,n}\colon \Delta(U_i \times H_{-i,n-1}) \times H_{-i,n-1} \to \Delta(Z)$ as in (1). By Lemma 1, there exist $0 < \varepsilon_0 \leq \varepsilon_1 \leq \cdots \leq \varepsilon_{N-1} \leq \varepsilon_N \leq \bar{\varepsilon}/2$ such that if $d_{i,n}(t_{i,n}, m_{i,n}) \leq \varepsilon_{n-1}$,

$d_{i,n}(t_{i,n}, m'_{i,n}) > \varepsilon_n$, and $|O_{i,n}(\cdot \mid \cdot, \cdot, \omega) - O^0_{i,n}| \leq \varepsilon_{n-1}$ for any $\omega \in \Omega$, then any preference in $\Delta_{\varepsilon_{n-1}, t_{i,n}}(U_i \times H_{-i,n-1} \times \Omega)$ strictly prefers $O_{i,n}(\cdot \mid m_{i,n}, \cdot, \cdot)$ to $O_{i,n}(\cdot \mid m'_{i,n}, \cdot, \cdot)$.

We define a mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$ as follows. For each $i \in I$ and $n \leq N$, let $M_{i,n}$ be any ε_{n-1}-dense finite subset of $\Delta(U_i \times H_{-i,n-1})$ with respect to $d_{i,n}$, and $M_i = \prod_{n=1}^N M_{i,n}$. Define $O \colon M \to \Delta(Z)$ by

$$O(z \mid m) = \frac{1-\delta}{|I|(1-\delta^N)} \sum_{i \in I} \sum_{n=1}^{N} \delta^{n-1} O^0_{i,n}(z \mid m_{i,n}, m_{-i,1}, \ldots, m_{-i,n-1})$$

for each $m \in M$ and $z \in Z$, where $\delta > 0$ is small enough to satisfy $(1-\delta)/\delta \geq (|I|-1)(1-\varepsilon_0)/\varepsilon_0$.

Claim 1 *For any type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$, any agent $i \in I$, and any type $t_i \in T_i$, we have*

$$m_i \in R_i^n(t_i; \mathcal{T}, \mathcal{M}) \Rightarrow d_{i,n}(\hat{\mu}_{i,n}(t_i), m_{i,n}) \leq \varepsilon_n$$

for any $n \leq N$.

We can now complete the proof of Proposition 2.

Proof of Proposition 2. Pick any pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$, $i \in I$, $t_i \in T_i$, and $t'_i \in T'_i$. Suppose that there exists $m_i = (m_{i,1}, \ldots, m_{i,N}) \in R_i(t_i; \mathcal{T}, \mathcal{M}) \cap R_i(t'_i; \mathcal{T}', \mathcal{M})$. 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

$$d_{i,n}(\hat{\mu}_{i,n}(t_i; \mathcal{T}), \hat{\mu}_{i,n}(t'_i; \mathcal{T}')) \leq d_{i,n}(\hat{\mu}_{i,n}(t_i; \mathcal{T}), m_{i,n}) + d_{i,n}(\hat{\mu}_{i,n}(t'_i; \mathcal{T}'), m_{i,n}) \leq 2\varepsilon_n \leq \bar{\varepsilon}$$

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

3.6 Equilibrium and Strategic Distinguishability

Our analysis thus far concerned the solution concept of interim correlated rationalizability. Our focus in this paper is on equilibrium. 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. We say that a profile $\sigma = (\sigma_i)_{i \in I}$ of measurable mappings $\sigma_i \colon \widehat{T}_i \to \Delta(M_i)$ is a *local equilibrium* of $(\mathcal{T}, \mathcal{M})$ on a belief closed subspace $\widehat{\mathcal{T}} = (\widehat{T}_i, \mu_i|_{\widehat{T}_i})_{i \in I}$ if

$$\int_{U_i \times T_{-i}} \sum_{m_{-i} \in M_{-i}} \sum_{z \in Z} u_i(z)(O(z \mid m_i, m_{-i}) - O(z \mid m'_i, m_{-i})) \left(\prod_{j \neq i} \sigma_j(m_j \mid t_j) \right) \mu_i(du_i, dt_{-i} \mid t_i) \geq 0$$

for any agent i, any $t_i \in \widehat{T}_i$, and any $m_i, m'_i \in M_i$ with $\sigma_i(m_i \mid t_i) > 0$. We denote by $E_i(t_i)$ the set of messages played with positive probability in local equilibrium on a belief closed subspace containing t_i. We also write $E_i(t_i; \mathcal{T}, \mathcal{M})$.

Proposition 3 *For any type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$, any agent $i \in I$, and any type $t_i \in T_i$, we have*

$$E_i(t_i; \mathcal{T}, \mathcal{M}) \supseteq E_i(\hat{\mu}_i(t_i); \mathcal{T}^*, \mathcal{M})$$

for any mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$.

Proof. The proposition follows because if $\sigma^* = (\sigma_i^*)_{i \in I}$ is a local equilibrium of $(\mathcal{T}^*, \mathcal{M})$, then by the belief-preserving property of $\hat{\mu}_i$, $\sigma = (\sigma_i)_{i \in I}$ with $\sigma_i = \sigma_i^* \circ \hat{\mu}_i$ is a local equilibrium of $(\mathcal{T}, \mathcal{M})$. ■

In general, equilibrium exists only under much stronger conditions that interim correlated rationalizability. But a sufficient condition for existence is that each T_i is countable and M_i is finite (a maintained assumption of the paper). This gives us:

Theorem 2 *For any two countable type spaces $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ and $\mathcal{T}' = (T_i', \mu_i')_{i \in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t_i' \in T_i'$, the following two conditions are equivalent:*

1. *$\hat{\mu}_i(t_i; \mathcal{T}) = \hat{\mu}_i(t_i'; \mathcal{T}')$;*
2. *$E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$ for any mechanism $\mathcal{M}$.*

Proof. 1 $\Rightarrow$ 2 follows from Proposition 3 and the existence of equilibria; 2 $\Rightarrow$ 1 follows from Theorem 1 (indeed Proposition 2) and the fact that local equilibrium is a refinement of rationalizability. ■

It is immediate to extend Theorem 2 to other interim solution concepts that are nonempty, coarser than equilibrium, and finer than interim correlated rationalizability. Strategic equivalence, however, holds only for interim correlated rationalizability and not for equilibrium. It is well known from environments with common certainty of preferences that solution concepts such as equilibrium and interim independent rationalizability depend on redundant types. See Dekel, Fudenberg, and Morris (2007), Ely and Peski (2006) and Sadzik (2010).

4 Discussion

4.1 The Expected Utility Assumption

We maintained the assumption of expected utility maximization. Epstein and Wang (1996) construct a universal type space of non-expected utility preferences, incorporating non-expected utility preferences such as ambiguity aversion, but maintaining monotonicity as well as additional regularity conditions. The universal type space we study when we relax the simplex assumption in

the next Section is expressed in a similar language to Epstein and Wang (1996). Unlike Epstein and Wang (1996), however, we impose independence to get an expected utility representation, and dispense with monotonicity to incorporate the interdependence of preferences we want to capture. di Tillio (2008) allows general preferences, and thus does not require monotonicity or independence, but restricts attention to preferences over finite outcomes at every order of the hierarchy.[15]

4.2 Separating Beliefs from Payoffs

Our approach integrates the treatment of payoffs (or ex post preferences) with beliefs and higher order beliefs about those payoffs. But the standard approach in the literature has been to discuss the two separately. In particular, as we discussed in Section 2, an alternative construction is to first identify a universal space of interdependent payoff types (with a single "characteristic"), following Gul and Pesendorfer (2010), and then allow for all possible beliefs and higher order beliefs over those payoff types. But this construction gives a different space from our universal space. In particular, it is neither coarser nor finer than our construction. The alternative construction makes distinctions which our space does not because, as discussed in Section 2, it will separate types that differ only in what their ex post preferences would be conditional on zero probability events. On the other hand, our construction allows ex post preferences to vary depending on others' beliefs. Thus I might be more altruistic if I believe that you believe that I am altruistic. This cannot arise in the alternative construction, where payoff types depend only on others' payoff types, not their beliefs. Thus we allow ex post preferences to depend on beliefs, as in the "psychological games" literature of Geanakoplos, Pearce, and Stacchetti (1989) and Battigalli and Dufwenberg (2009). However, preferences depend only on other agents' beliefs about others' types and not - as in the psychological games literature - on beliefs about actions.[16]

However, even though our universal space is quite different from this alternative construction, it still makes sense to ask if and how we can distinguish between "payoff types" and "belief types" in a natural way in our universal space. Just as beliefs cannot be pinned down in (single person) expected utility representations of preferences unless we fix a numéraire, there is indeterminacy in beliefs in our construction based on the choice of representations of the extreme preferences U_i. But particular applications may suggest a numéraire over which the modeler wishes to treat utility

[15] The strategic distinguishability question does not appear to have been addressed without expected utility preferences. Chambers (2008) shows the impossibility of constructing a uniform scoring rule to distinguish preferences and beliefs in a non-expected utility setting, which suggests that positive results about strategic distinguishability would be hard to obtain. Grant, Meneghel, and Tourky (2014) analyze "Savage games" played on subjective state spaces, allowing both expected utility maximizers and more general preferences; in both cases, they do not construct a universal type space or consider strategic distinguishability.

[16] Such beliefs can be captured as "characteristics" in Gul and Pesendorfer (2010).

as state independent, which will pin down the representation. A universal type of agent i can then be characterized by a belief over others' types, and ex post preferences over outcomes conditional on others' types. We can use the separation to interpret existing works.

Dasgupta and Maskin (2000) consider mechanisms where agents report what their valuations would be conditional on others' valuations taking particular values. This is a subset of the information expressed in our universal type space. Under their maintained assumptions about what is common knowledge among the agents, this is enough information to identify true valuations. Bergemann and Morris (2009) fix a payoff type space and show that payoff types can be strategically distinguished without reference to their beliefs and higher order beliefs if and only if there is not "too much" interdependence in ex post preferences.

4.3 Strategic Revealed Preference

We have identified an operational definition of types. Two types are strategically distinguishable if there exists a mechanism where they are guaranteed to behave differently. This definition is operational, and does not make reference to information or signals that agents have observed. We do not consider whether agents' interdependent types' preferences are based on information because we do not know what the operational counterpart to that information is.[17]

Classical single person revealed preference theory characterizes when a set of choice functions are consistent with rational choice (Afriat (1967)), with the weak axiom of revealed preference (WARP) being the key restriction on choice rules. If, in additional to standard rationality assumptions, we looked at choices over lotteries and added the independence assumption, we would obtain more restrictions. A primitive single person revealed preference question would then be if you can tell the difference between two different expected utility preferences over lotteries. A standard argument says that we can construct a pair of lotteries such that one preference will lead to one strict ordering, and the other preference will lead to the opposite strict ordering. Our strategic distinguishability question is a many person analogue of this question.[18]

[17] This contrasts with one approach in the epistemic game theory literature that argues that it is necessary to include an explicit description of signals to analyze environments with asymmetric information. See, for example, Battigalli, Tillio, Grillo, and Penta (2011).

[18] There is a small literature developing strategic analogues of classic single agent decision theory. See, for example, Sprumont (2000). There are many differences between this paper and that literature. Thus Sprumont (2000) fix the action set while we allow for arbitrary action sets. Their theory is ordinal and does not impose expected utility while ours is cardinal and does impose expected utility.

4.4 Incorporating External States

We made a modelling choice to restrict attention to "uncontingent mechanisms," $O\colon M \to \Delta(Z)$, where agents' actions alone determined outcomes. We modelled agents' interdependent preferences, which entailed modelling the agents' incomplete information about each others' preferences. The maintained interpretation was that this incomplete information was about preferences which, mathematically, could be represented as beliefs over private states corresponding to extreme points of their own ex post preferences.

However, game theorists often talk about incomplete information about external and verifiable states which may influence which outcomes the mechanism will choose. If we write Θ for a finite (or compact metric) space of external states that influence outcomes, then we can consider a richer class of contingent mechanisms, $O\colon M \times \Theta \to \Delta(Z)$. With a richer class of mechanisms, we will be able to more finely strategically distinguish types, since these external states may also impact preferences and beliefs and higher order beliefs about them may also be revealed. We excluded discussion of such external states earlier because they were incidental to our primary exercise of characterizing the universal type space of interdependent preferences. But reporting this extension now allows us to connect our result to those of DFM, according to their original interpretation, in an exact way (see the next sub-Section). The results and proofs do not change, once we alter our definitions of type spaces, mechanisms and solution concepts to reflect Θ in the appropriate way. Thus we will merely state how the definitions must be changed in order for our previous results to hold as stated.

A type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ now consists of non-empty measurable spaces T_i of agent i's possible types and measurable mappings $\mu_i\colon T_i \to \Delta(U_i \times \Theta \times T_{-i})$, i.e., a belief type space over private states and external states $(U_i \times \Theta)_{i\in I}$. The universal type space $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i\in I}$ is constructed with the homeomorphism $\mu_i^*\colon T_i^* \to \Delta(U_i \times \Theta \times T_{-i}^*)$. For any type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$, the mapping $\hat{\mu}_i\colon T_i \to T_i^*$ maps each type in T_i to its hierarchy of beliefs over $(U_i \times \Theta)_{i\in I}$. A mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$ consists of non-empty finite sets M_i of messages available to agent i and the outcome function $O\colon M \times \Theta \to \Delta(Z)$. The definition of rationalizability becomes:

$$
\begin{aligned}
R_i^0(t_i) &= M_i, \\
R_i^{n+1}(t_i) &= \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists a measurable mapping } \sigma_{-i}\colon U_i \times \Theta \times T_{-i} \to \Delta(M_{-i}) \text{ s.t.:} \\ \text{(i) } \sigma_{-i}(R_{-i}^n(t_{-i}) \mid u_i, \theta, t_{-i}) = 1 \text{ for any } u_i \in U_i,\ \theta \in \Theta \text{ and } t_{-i} \in T_{-i}, \\ \text{(ii) } \int_{U_i\times\Theta\times T_{-i}} \sum_{m_{-i}\in M_{-i}} \sum_{z\in Z} u_i(z)(O(z \mid m_i, m_{-i}, \theta) - O(z \mid m_i', m_{-i}, \theta)) \\ \qquad \times \sigma_{-i}(m_{-i} \mid u_i, \theta, t_{-i})\mu_i(du_i, d\theta, dt_{-i} \mid t_i) \geq 0, \text{ for any } m_i' \in M_i. \end{array} \right\}, \\
R_i(t_i) &= \bigcap_{n=0}^{\infty} R_i^n(t_i).
\end{aligned}
$$

We say that a profile $\sigma = (\sigma_i)_{i \in I}$ of measurable mappings $\sigma_i \colon \widehat{T}_i \to \Delta(M_i)$ is a local equilibrium on a belief closed subspace $\widehat{\mathcal{T}} = (\widehat{T}_i, \mu_i|_{\widehat{T}_i})_{i \in I}$ if

$$\int_{U_i \times \Theta \times T_{-i}} \sum_{m_{-i} \in M_{-i}} \sum_{z \in Z} u_i(z)(O(z \mid m_i, m_{-i}, \theta) - O(z \mid m_i', m_{-i}, \theta)) \left(\prod_{j \neq i} \sigma_j(m_j \mid t_j) \right) \mu_i(du_i, d\theta, dt_{-i} \mid t_i) \geq 0$$

for any agent i, any $t_i \in \widehat{T}_i$, and any $m_i, m_i' \in M_i$ with $\sigma_i(m_i \mid t_i) > 0$.

Now Theorems 1 and 2 remain true as stated. Our previous analysis corresponds to the special case where Θ is a singleton.

4.5 Common Certainty of Preferences

We now argue that if we extend our framework to incorporate external states (as in the previous sub-Section), but impose the restriction that there is common certainty of ex post preferences, then we map our problem and results back to the environment studied by Dekel, Fudenberg, and Morris (2006), (2007).[19]

Say that there is *common certainty of preferences* if each U_i is a singleton $\{u_i\}$, where $u_i \in \mathbb{R}^Z$ is not constant over Z. This is an example of a simplex restriction as defined in Section 3.1.[20] Under common certainty of preferences, there is uncertainty and higher order uncertainty about external states but no uncertainty about preferences. Thus the universal type space is simply the Mertens-Zamir universal type space, corresponding to the set of coherent belief hierarchies about external states Θ.

Given that each U_i is a singleton, picking a contingent mechanism is equivalent to picking a game (a specification of payoffs as a function of message/action profiles and external states), with the proviso that the set of feasible payoff vectors is given by the convex hull of the set of payoff vectors that can arise from a given outcome. Write V for the set of payoff profiles that can be induced by some lottery over outcomes, so that

$$V = \text{conv}\{(u_i(z))_{i \in I} \in \mathbb{R}^I \mid z \in Z\}.$$

Now consider a game $\mathcal{G} = ((M_i)_{i \in I}, g)$, where M_i is the set of actions for agent i,

$$g \colon M \times \Theta \to V,$$

[19] There is an alternative interpretation of DFM under which they can be seen as a special case of the results in this paper without appeal to "external" states. Observe that uncontingent mechanisms and states - profiles of extremal preferences, in our simplex formulation - jointly define a set of utility functions from message profiles and states to payoffs, i.e., a game. If the outcome space were sufficiently rich, this problem would reduce to DFM. If not, results in this paper would identify strategic distinguishability in restricted classes of games.

[20] Property 2 in Section 3.1 is maintained and the other properties are vacuous in this case.

and $g_i(m,\theta)$ is the payoff of agent i if action profile m is chosen and the external state is θ. Call such a game a *V-game*. Each contingent mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$ induces a V-game $\mathcal{G} = ((M_i)_{i\in I}, g)$ with

$$g(m,\theta) = \left(\sum_z u_i(z)O(z \mid m,\theta)\right)_{i\in I}$$

for any $m \in M$ and $\theta \in \Theta$; conversely, any V-game can be induced by some contingent mechanism.

Our definition of interim correlated rationalizability in this case corresponds exactly to that in Dekel, Fudenberg, and Morris (2006), (2007). Our Theorem 1 now proves that two types have the same belief hierarchy over Θ if and only if they have the same ICR actions in all V-games. In the case that V is a non-degenerate product set, i.e.,

$$V = \prod_{i\in I} [\underline{v}_i, \overline{v}_i]$$

with $\underline{v}_i < \overline{v}_i$ for any $i \in I$, this result was already proved in Dekel, Fudenberg, and Morris (2006), (2007). Specifically, for any non-degenerate product set V, Dekel, Fudenberg, and Morris (2006, Lemma 4) show that if two types have distinct belief hierarchies, then there is a V-game where they have disjoint rationalizable action sets;[21] conversely, Dekel, Fudenberg, and Morris (2007, Proposition 1 and Corollary 2) show that two types with the same belief hierarchy have the same ICR actions (for finite types and general types, respectively) in any V-game.

The assumption that the set V is a non-degenerate product set has a natural counterpart in our setting. Say that we have a *private good environment* if the outcome space Z has a product structure $Z = \prod_{i\in I} Z_i$, and each agent i's utility from outcome z depends only on the ith component z_i, so $u_i(z) = \widetilde{u}_i(z_i)$ for some $\widetilde{u}_i : Z_i \to \mathbb{R}$. In this case, the set of feasible payoff vectors has the product structure

$$V = \prod_{i\in I} [\underline{v}_i, \overline{v}_i],$$

where

$$\underline{v}_i = \min_{z_i\in Z_i} \widetilde{u}_i(z_i) \text{ and } \overline{v}_i = \max_{z_i\in Z_i} \widetilde{u}_i(z_i).$$

But our Theorem 1 did not rely on the private good environment assumption. If the common certainty of preferences assumption is maintained but the private good assumption is dropped, then the set V of feasible payoff profiles could be any convex polytope whose projection in any

[21] Dekel, Fudenberg, and Morris (2006, Lemma 4) prove something a little stronger: for any distance between nth order beliefs, we can find $\varepsilon > 0$ such that any action which is δ-rationalizable for one type is not even $(\delta+\varepsilon)$-rationalizable for the other type.

dimension is non-degenerate. For example, our Theorem would apply to environments where

$$V = \left\{ v \in [-1,1]^I \left| \sum_{i \in I} v_i = 0 \right. \right\}$$

so we restricted attention to zero sum games. And it would apply to environments where

$$V = \left\{ v \in [0,1]^I \left| v_i = v_j \text{ for all } i,j \in I \right. \right\},$$

so we restricted to common interest games. Thus, while the original proof of Dekel, Fudenberg, and Morris (2006, Lemma 4) relied on the assumption that any payoff vectors are feasible, our Theorem 1 - with external states added and common certainty of preferences assumed - establishes that it would remain true if DFM had restricted attention to zero sum games, common interest games, or many other subsets of games which restricted how agents' payoffs can vary.

Gossner and Mertens (2001) show that a zero sum game of incomplete information has a value which depends only on the probability distribution over Mertens-Zamir hierarchies and is increasing in informativeness in Blackwell's sense. The argument requires a strategic distinguishability result for the case of zero sum games.[22] While the formulation of our strategic distinguishability question and the proof are different from those arising in Gossner and Mertens (2001), the argument above suggests when and how the approach in this paper could be used to develop analogous strategic distinguishability exercises in different classes of games.

5 General Hierarchies of Preferences and λ-Continuity

We imposed a simplex restriction on agents' ex post preferences. We made this assumption because it not only sufficed for the compactness properties that we need for our main results, but also allowed us to state the results in as standard a language as possible.

The simplex restriction is a strong restriction in many ways. It does not have a clear interpretation based on preferences. It has the strong implication that there exists a pair of lotteries for each agent with a strict ranking that is independent of his own type as well as the opponents' types. This property was not assumed by Abreu and Matsushima (1992b) and thus our simplex results do not formally imply their Lemma 2.

Without the simplex assumption, it is not natural to represent preferences by beliefs over ex post preferences. Thus, in this Section, we introduce an alternative language for describing preferences and preference hierarchies more directly. We show that without any restriction on ex

[22] Gossner and Mertens (2001) is an abstract of unpublished work; we are grateful to Olivier Gossner for privately sharing notes from the complete paper.

post preferences, no mechanism can strategically distinguish two types with the same first order preference (i.e., the same preference over lotteries). Finally, we explore the restriction of what we call "λ-continuity," which is weaker than any simplex restriction, and yet sufficient for the strategic distinguishability result. Namely, we establish that as long as we restrict attention to λ-continuous types, two types are strategically distinguishable if and only if they are mapped to distinct preference hierarchies.

Note that we keep the discussion in this Section rather informal. We also restrict attention to countable type spaces and to the solution concept of equilibrium. A more formal treatment and proofs are presented in Supplemental Appendix, which also shows how the results can be extended to general type spaces and more permissive rationalizability solution concepts.

5.1 No Restriction on Ex Post Preferences

We relax the simplex restriction, and allow for a more general class of interdependent expected utility preferences. For simplicity, we consider a countable type space represented by $\mathcal{T} = (T_i, \mu_i, u_i)_{i \in I}$ with

$$\mu_i \colon T_i \to \Delta(T_{-i}),$$
$$u_i \colon T_i \times T_{-i} \times Z \to \mathbb{R},$$

where for each $t_i \in T_i$, $u_i(\cdot \mid t_i, \cdot)$ is absolutely summable with respect to $\mu_i(\cdot \mid t_i)$, i.e., $\sum_{t_{-i}, z} |u_i(z \mid t_i, t_{-i})| \mu_i(t_{-i} \mid t_i) < \infty$. Given a mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$ with $O \colon M \to \Delta(Z)$, a profile $\sigma = (\sigma_i)_{i \in I}$ of behavioral strategies $\sigma_i \colon T_i \to \Delta(M_i)$ is an *equilibrium* if

$$\sum_{t_{-i} \in T_{-i}} \sum_{m_{-i} \in M_{-i}} \sum_{z \in Z} u_i(z \mid t_i, t_{-i}) (O(z \mid m_i, m_{-i}) - O(z \mid m_i', m_{-i})) \left(\prod_{j \neq i} \sigma_j(m_j \mid t_j) \right) \mu_i(t_{-i} \mid t_i) \geq 0$$

for any agent $i \in I$, any type $t_i \in T_i$, and any (truthful and deviating) messages $m_i, m_i' \in M_i$ with $\sigma_i(m_i \mid t_i) > 0$.

The above type space $\mathcal{T} = (T_i, \mu_i, u_i)_{i \in I}$ is based on belief-utility representations, and unlike in the case of simplex restriction, each preference has multiple belief-utility representations, all of which are equally unnatural. A more natural approach is to work directly with preferences as follows. For a countable set X, recall that $F(X)$ denotes the set of all (Anscombe-Aumann) acts over X. Let $P(X)$ be the set of all preferences $\succsim$ over $F(X)$ represented by $\mu \in \Delta(X)$ and μ-absolutely summable state-dependent utility $u \colon X \times Z \to \mathbb{R}$ as follows:

$$f \succsim f' \Leftrightarrow \sum_{x \in X} \sum_{z \in Z} u(z \mid x)(f(z \mid x) - f'(z \mid x)) \mu(x) \geq 0.$$

With this P-notation, a type space is simply expressed as $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ with

$$\pi_i \colon T_i \to P(T_{-i}),$$

where $\pi_i(t_i)$ is the preference of type t_i over acts over the opponents' types. This new language is rich enough to describe all necessary information about the type space, including the notion of equilibrium without explicit references to belief-utility representations. We could have expressed our earlier simplex results in this language also, but chose not to do so in order to highlight the tight connection with the existing literature.[23]

Given a mechanism $\mathcal{M}$, a behavioral strategy profile σ is an *equilibrium* if $\pi_i(t_i)$ weakly prefers $O(\cdot \mid m_i, \cdot) \circ \sigma_{-i}$ to $O(\cdot \mid m_i', \cdot) \circ \sigma_{-i}$ for any i, t_i, m_i, m_i' with $\sigma_i(m_i \mid t_i) > 0$.[24] As before, let $E_i(t_i; \mathcal{T}, \mathcal{M})$ denote the set of all messages that can be played with positive probability in equilibrium. Given a type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ with $\pi_i \colon T_i \to P(T_{-i})$, we write $\hat{\pi}_{i,1}(t_i)$ for the unconditional preference of type t_i over lotteries, and call it his first order preference. Note that we have $\hat{\pi}_{i,1}(t_i) \in P(\{*\})$, where $\{*\}$ denotes an arbitrary singleton set. We also write $\hat{\pi}_{i,2}(t_i)$ for the restriction of type t_i's preference over acts that depend only on the opponents' first order preferences, i.e., for any f, f', $\hat{\pi}_{i,2}(t_i)$ weakly prefers f to f' if and only if $\pi_i(t_i)$ weakly prefers $f \circ \hat{\pi}_{-i,1}$ to $f' \circ \hat{\pi}_{-i,1}$, and we call it his second order preference;[25] thus $\hat{\pi}_{i,2}(t_i) \in P((P(\{*\}))^{|I|-1})$. We define third order, and higher order, preferences similarly, and we write $\hat{\pi}_i(t_i) = (\hat{\pi}_{i,1}(t_i), \hat{\pi}_{i,2}(t_i), \ldots)$ for the hierarchy of type t_i's higher order preferences. Let H_f be the set of all preference hierarchies of finite types.

Now the following result will continue to be true.

Proposition 4 *For any two countable type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T_i', \pi_i')_{i \in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t_i' \in T_i'$, if $\hat{\pi}_i(t_i; \mathcal{T}) = \hat{\pi}_i(t_i'; \mathcal{T}')$, then $E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$ for any mechanism $\mathcal{M}$.*

However, the converse will not be true. In particular, we have:

[23] In an earlier version of this work, Bergemann, Morris, and Takahashi (2011), we expressed simplex results (implied by a "worst outcome assumption") in this preference language.

[24] Note that we have $O(\cdot \mid m_i, \cdot) \circ \sigma_{-i} \in F(T_{-i})$ defined by

$$(O(\cdot \mid m_i, \cdot) \circ \sigma_{-i})(z \mid t_{-i}) = \sum_{m_{-i} \in M_{-i}} O(z \mid m_i, m_{-i}) \prod_{j \neq i} \sigma_j(m_j \mid t_j)$$

for each $t_{-i} \in T_{-i}$ and $z \in Z$. Similarly, we have $O(\cdot \mid m_i', \cdot) \circ \sigma_{-i} \in F(T_{-i})$.

[25] Note that we have $f \circ \hat{\pi}_{-i,1} \in F(T_{-i})$ defined by

$$(f \circ \hat{\pi}_{-i,1})(z \mid t_{-i}) = f\left(z \mid (\hat{\pi}_{j,1}(t_j))_{j \neq i}\right)$$

for each $t_{-i} \in T_{-i}$ and $z \in Z$. Similarly, we have $f' \circ \hat{\pi}_{-i,1} \in F(T_{-i})$.

Proposition 5 *Suppose that two preference hierarchies of finite types* $h = (\succsim_1, \succsim_2, \ldots), h' = (\succsim'_1, \succsim'_2, \ldots) \in H_f$ *share the same first order preference, i.e.,* $\succsim_1 = \succsim'_1$. *Then for any agent* $i \in I$ *and any mechanism* $\mathcal{M} = ((M_i)_{i\in I}, O)$, *there exist two finite type spaces* $\mathcal{T} = (T_i, \pi_i)_{i\in I}$ *and* $\mathcal{T}' = (T'_i, \pi'_i)_{i\in I}$ *and two types* $t_i \in T_i$ *and* $t'_i \in T'_i$ *such that* $\hat{\pi}_i(t_i; \mathcal{T}) = h$, $\hat{\pi}_i(t'_i; \mathcal{T}') = h'$, *and* $E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t'_i; \mathcal{T}', \mathcal{M}) \neq \emptyset$.

The proof is in the Supplemental Appendix. In what follows, we use an example to illustrate why a mechanism like the one constructed in Section 3.5.1 cannot strategically distinguish types with distinct hierarchies of preferences.

Take the conditional altruism scenario, and consider a mechanism with two messages 0 and 1 for each agent and the outcome being in the form of

$$O(\cdot \mid m_1, m_2) = (1-\varepsilon)O_1(\cdot \mid m_1) + \varepsilon O_2(\cdot \mid m_1, m_2)$$

with $\varepsilon \geq 0$, where O_1 is to solicit agent 1's report about his first order preference, whereas O_2 is to solicit both agents' reports about their higher order preferences. To fix ideas, suppose that $O_1(\cdot \mid m_1 = 0)$ gives the prize to nobody and agent 1 with probability $\frac{1}{2}$ each, and $O_1(\cdot \mid m_1 = 1)$ gives the prize to agent 2; $O_2(\cdot \mid m_1, m_2)$ gives the prize to agents 1 and 2 with probability $\frac{m_1 m_2}{2}$, and to nobody with the remaining probability $1 - m_1 m_2$. Consider a type space, where each agent i has two possible types 0 and 1, each type believes that the opponent's type is 0 or 1 with probability $\frac{1}{2}$, and payoff parameters (the payoff from the opponent getting the prize) are given by

	$t_2 = 0$	1
$t_1 = 0$	$1+v, 1+v$	$1-v, 1$
1	$1, 1-v$	$1, 1$

with $v \in \mathbb{R}$. Note that all types have the same expected value of the payoff parameter $\frac{1+v}{2} + \frac{1-v}{2} = 1$, and hence have the same preference hierarchy as the truly altruistic type with complete information, independently of v.

In this case, if $\varepsilon = 0$, then agent 1 has an incentive to report $m_1 = 1$ (as a dominant action) according to his first order preference. But since there is no interaction term between m_1 and m_2, no information about higher order preferences can be revealed in equilibrium actions. In contrast, if $\varepsilon > 0$, then for sufficiently large v, type 0 of agent 1 no longer has a dominant action, and indeed, the strategy profile of reporting $m_i = t_i$ becomes an equilibrium. In sum, there is no $\varepsilon \geq 0$ that keeps agent 1's incentive to report his first order preference truthfully and yet solicits higher order preferences from any agent.

Note that this example hinges crucially on the difference between us and AM: our exercise of strategic distinguishability (Theorems 1 and 2) is to construct a mechanism independently of an

underlying type space, whereas AM fix a finite type space first and then construct a mechanism. This example also illustrates a trade-off between ε and v. There is no $\varepsilon > 0$ that keeps $m_1 = 1$ a dominant action for agent 1 independently of v. But if we knew a bound of v, then we could choose ε small enough (in the magnitude of $1/|v|$) so that $m_1 = 1$ is a dominant action for agent 1. The notion of λ-continuity in the next sub-Section formalizes this idea in a general type space.

5.2 λ-Continuity

For a given $\lambda > 0$, we say that a preference $\succsim \in P(X)$ is λ-*continuous* if there exists a pair of outcomes z and z' such that $z \succ z'$ and for any $f, f' \in F(X)$,

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

That is, we require that the preference relation $z \succ z'$ be maintained at least weakly even if we mix these uncontingent outcomes z and z' with a small probability of state-contingent lotteries (i.e., acts) f and f'. In other words, λ-continuity imposes a bound on the "intensity" of state dependency measured by the utility difference between any two acts relative to the utility difference between any two outcomes. Let $P_\lambda(X)$ be the set of all λ-continuous preferences in $P(X)$. $P_\lambda(X)$ for small $\lambda > 0$ can be thought of all preferences excluding a neighborhood of preferences that are completely indifferent over all outcomes. A type is said to be λ-continuous if it belongs to a type space where all types are λ-continuous.

We can now prove a converse to Proposition 4 for λ-continuous preferences:

Proposition 6 *Fix $\lambda > 0$. For any two λ-continuous type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T_i', \pi_i')_{i \in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t_i' \in T_i'$, if $\hat{\pi}_i(t_i; \mathcal{T}) \neq \hat{\pi}_i(t_i'; \mathcal{T}')$, then there exists $\mathcal{M}$ with $E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t_i'; \mathcal{T}', \mathcal{M}) = \emptyset$.*

Thus Propositions 4 and 6 together show a version of Theorem 2 for countable and λ-continuous preferences.[26] Lemma 2 of Abreu and Matsushima (1992b) is a special case of Proposition 6, since finite type spaces are λ-continuous.

[26] In the Supplemental Appendix, we will construct a universal type space of interdependent preferences without a topological structure, analogous to the topology free construction of the universal type space in Heifetz and Samet (1998). However, one can also follow the classical construction of the universal type space like in Mertens and Zamir (1985) by exploiting the compact metrizability of $P_\lambda(X)$.

A Appendix

Proof of Proposition 1. We will show

$$\forall i \in I,\ \forall t_i \in T_i, \quad R_i^n(t_i;\mathcal{T},\mathcal{M}) = R_i^n(\hat{\mu}_i(t_i);\mathcal{T}^*,\mathcal{M}) \qquad (\text{Equiv}_n)$$

by induction on n.

It is obvious that (Equiv$_0$) holds.

Suppose that (Equiv$_n$) holds for $n \geq 0$. Fix any $i \in I$ and any $t_i \in T_i$. For any $m_i \in R_i^{n+1}(t_i;\mathcal{T},\mathcal{M})$, there exists $\nu_i \in \Delta(M_{-i}\times U_i\times T_{-i})$ such that $\nu_i\{(m_{-i},u_i,t_{-i}) \mid m_{-i} \in R_{-i}^n(t_{-i};\mathcal{T},\mathcal{M})\} = 1$, $\mathrm{mrg}_{U_i\times T_{-i}}\,\nu_i = \mu_i(t_i)$, and $\int_{M_{-i}\times U_i\times T_{-i}} \sum_{z\in Z} u_i(z)(O(z \mid m_i,m_{-i}) - O(z \mid m_i',m_{-i}))\nu_i(dm_{-i},du_i,dt_{-i}) \geq 0$ for any $m_i' \in M_i$. Let $\nu_i^* \in \Delta(M_{-i}\times U_i\times T_{-i}^*)$ be such that

$$\nu_i^*(E) = \nu_i(\{(m_{-i},u_i,t_{-i}) \mid (m_{-i},u_i,\hat{\mu}_{-i}(t_{-i})) \in E\})$$

for any measurable subset $E \subseteq M_{-i}\times U_i\times T_{-i}$. Then we have

$$\begin{aligned} & \nu_i^*\{(m_{-i},u_i,t_{-i}^*) \mid m_{-i} \in R_{-i}^n(\hat{\mu}_{-i}(t_{-i});\mathcal{T}^*,\mathcal{M})\} \\ = \ & \nu_i^*\{(m_{-i},u_i,t_{-i}^*) \mid m_{-i} \in R_{-i}^n(t_{-i};\mathcal{T},\mathcal{M})\} = 1 \end{aligned}$$

by the induction hypothesis, $\mathrm{mrg}_{U_i\times T_{-i}^*}\,\nu_i^* = \mu_i^*(\hat{\mu}_i(t_i))$ by the belief-preserving property of $\hat{\mu}_i$, and

$$\begin{aligned} & \int_{M_{-i}\times U_i\times T_{-i}^*} \sum_{z\in Z} u_i(z)(O(z \mid m_i,m_{-i}) - O(z \mid m_i',m_{-i}))\nu_i^*(dm_{-i},du_i,dt_{-i}^*) \\ = & \int_{M_{-i}\times U_i\times T_{-i}} \sum_{z\in Z} u_i(z)(O(z \mid m_i,m_{-i}) - O(z \mid m_i',m_{-i}))\nu_i(dm_{-i},du_i,dt_{-i}) \geq 0 \end{aligned}$$

for any $m_i' \in M_i$. Thus $m_i \in R_i^{n+1}(\hat{\mu}_i(t_i);\mathcal{T}^*,\mathcal{M})$. Since this holds for any $m_i \in R_i^{n+1}(t_i;\mathcal{T},\mathcal{M})$, we have $R_i^{n+1}(t_i;\mathcal{T},\mathcal{M}) \subseteq R_i^{n+1}(\hat{\mu}_i(t_i);\mathcal{T}^*,\mathcal{M})$.

Conversely, for any $m_i \in R_i^{n+1}(\hat{\mu}_i(t_i);\mathcal{T}^*,\mathcal{M})$, there exists a measurable mapping $\sigma_{-i}^*\colon U_i\times T_{-i}^* \to \Delta(M_{-i})$ such that $\sigma_{-i}^*(R_{-i}^n(\hat{\mu}_{-i}(t_{-i});\mathcal{T}^*,\mathcal{M}) \mid u_i,t_{-i}^*) = 1$ and $\int_{U_i\times T_{-i}^*} \sum_{z\in Z} u_i(z)(O(z \mid m_i,\sigma_{-i}^*(\cdot \mid u_i,t_{-i}^*)) - O(z \mid m_i',\sigma_{-i}^*(\cdot \mid u_i,t_{-i}^*)))\mu_i^*(du_i,dt_{-i}^* \mid \hat{\mu}_i(t_i)) \geq 0$ for any $m_i' \in M_i$. Let $\sigma_{-i}\colon U_i\times T_{-i} \to \Delta(M_{-i})$ be such that

$$\sigma_{-i}(m_{-i} \mid u_i,t_{-i}) = \sigma_{-i}^*(m_{-i} \mid u_i,\hat{\mu}_{-i}(t_{-i}))$$

for any $m_{-i} \in M_{-i}$, any $u_i \in U_i$, and any $t_{-i} \in T_{-i}$. Then we have $\sigma_{-i}(R_{-i}^n(t_{-i};\mathcal{T},\mathcal{M}) \mid u_i,t_{-i}) = \sigma_{-i}^*(R_{-i}^n(\hat{\mu}_{-i}(t_{-i});\mathcal{T}^*,\mathcal{M}) \mid u_i,\hat{\mu}_{-i}(t_{-i})) = 1$. Also, by the belief-preserving property of $\hat{\mu}_i$, we have

$$\begin{aligned} & \int_{U_i\times T_{-i}} \sum_{m_{-i}\in M_{-i}} \sum_{z\in Z} u_i(z)(O(z \mid m_i,m_{-i}) - O(z \mid m_i',m_{-i}))\sigma_{-i}(m_{-i} \mid u_i,t_{-i})\mu_i(du_i,dt_{-i} \mid t_i) \\ = & \int_{U_i\times T_{-i}^*} \sum_{m_{-i}\in M_{-i}} \sum_{z\in Z} u_i(z)(O(z \mid m_i,m_{-i}) - O(z \mid m_i',m_{-i}))\sigma_{-i}^*(m_{-i} \mid u_i,t_{-i}^*)\mu_i^*(du_i,dt_{-i}^* \mid \hat{\mu}_i(t_i)) \geq 0 \end{aligned}$$

for any $m_i' \in M_i$. Thus $m_i \in R_i^{n+1}(t_i; \mathcal{T}, \mathcal{M})$. Since this holds for any $m_i \in R_i^{n+1}(\hat{\mu}_i(t_i); \mathcal{T}^*, \mathcal{M})$, we have $R_i^{n+1}(t_i; \mathcal{T}, \mathcal{M}) \supseteq R_i^{n+1}(\hat{\mu}_i(t_i); \mathcal{T}^*, \mathcal{M})$.

Combining the both directions, we have (Equiv$_{n+1}$). ■

Proof of Lemma 1. Suppose not. Then there exists $\varepsilon > 0$ such that for every $n \in \mathbb{N}$, there exist $\mu_n, m_n, m_n' \in \Delta(U \times X)$ with $d_\Delta(\mu_n, m_n) \leq 1/n$ and $d_\Delta(\mu_n, m_n') > \varepsilon$, measurable space Ω_n, perturbed outcome function $O_n \colon M^0 \times X \times \Omega_n \to \Delta(Z)$ with $\|O_n(\cdot \mid \cdot, \cdot, \omega) - O^0\| \leq 1/n$ for every $\omega \in \Omega_n$, $\mu_n' \in \Delta(U \times X \times X \times \Omega_n)$ such that $\mu_n'(\{(u, x, x', \omega) \mid d(x, x') \leq \delta\}) = 1$, $\operatorname{mrg}_{1,2} \mu_n' = \mu_n$, and $\operatorname{mrg}_{1,3,4} \mu_n'$ weakly prefers $O_n(\cdot \mid m_n', \cdot, \cdot)$ to $O_n(\cdot \mid m_n, \cdot, \cdot)$. Since X is a compact metric space, by taking a subsequence if necessary, we can find $\mu^*, m'^{,*} \in \Delta(U \times X)$ such that $\mu_n \to \mu^*$ and $m_n' \to m'^{,*}$ as $n \to \infty$. Note that $m_n \to \mu^*$ as $n \to \infty$, and $\mu^* \neq m'^{,*}$. Let

$$u^* = \int \sum_z u(z) O^0(z \mid \mu^*, x) d\mu^*(u, x).$$

Claim 2 *We have*

$$\lim_{n \to \infty} \int \sum_z u(z) O^0(z \mid m_n, x) d\mu_n(u, x) = u^*,$$
$$\limsup_{n \to \infty} \int \sum_z u(z) O^0(z \mid m_n', x) d\mu_n(u, x) < u^*.$$

Proof of Claim 2. The claim follows from showing that

$$\lim_{n \to \infty} \int \sum_z u(z) C_{m_n}(f_k, f_l)(z \mid x) d\mu_n(u, x) = \int \sum_z u(z) C_{\mu^*}(f_k, f_l)(z \mid x) d\mu^*(u, x),$$
$$\limsup_{n \to \infty} \int \sum_z u(z) C_{m_n'}(f_k, f_l)(z \mid x) d\mu_n(u, x) \leq \int \sum_z u(z) C_{\mu^*}(f_k, f_l)(z \mid x) d\mu^*(u, x),$$

for each k, l, and that the second inequality holds with strict inequality for some k, l. The first equality and the second weak inequality follow from the standard revealed preference argument. To show the strict inequality, since $\mu^* \neq m'^{,*}$ and $F \subset F_c(X)$ is dense in the sup norm, there exist k, l such that μ^* strictly prefers f_k to f_l while $m'^{,*}$ strictly prefers f_l to f_k. Since m_n' strictly prefers f_l to f_k for sufficiently large n, we have:

$$\begin{aligned}
&\lim_{n \to \infty} \int \sum_z u(z) C_{m_n'}(f_k, f_l)(z \mid x) d\mu_n(u, x) \\
&= \lim_{n \to \infty} \int \sum_z u(z) f_l(z \mid x) d\mu_n(u, x) \\
&= \int \sum_z u(z) f_l(z \mid x) d\mu^*(u, x)
\end{aligned}$$

$$< \int \sum_z u(z) f_k(z \mid x) d\mu^*(u,x)$$
$$= \int \sum_z u(z) C_{\mu^*}(f_k, f_l)(z \mid x) d\mu^*(u,x).$$

which establishes the claim. ■

Claim 3 *We have*

$$\lim_{n\to\infty} \left(\int \sum_z u(z) O_n(z \mid m, x', \omega) d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) O^0(z \mid m, x) d\mu_n(u,x) \right) = 0$$

and the convergence is uniform in $m \in M^0$.

Proof of Claim 3. Note that

$$\left| \int \sum_z u(z) O_n(z \mid m, x', \omega) d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) O^0(z \mid m, x) d\mu_n(u,x) \right|$$
$$\leq \left| \int \sum_z u(z) O_n(z \mid m, x', \omega) d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) O^0(z \mid m, x') d\mu'_n(u,x,x',\omega) \right|$$
$$+ \left| \int \sum_z u(z) O^0(z \mid m, x') d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) O^0(z \mid m, x) d\mu_n(u,x) \right|.$$

The first term is bounded by $(1/n) \max_{u,z,z'} |u(z) - u(z')|$ since $\|O_n(\cdot \mid \cdot, \cdot, \omega) - O^0\| \leq 1/n$ for every $\omega \in \Omega_n$.

To show that the second term converges to 0 uniformly in m, it is enough to show that

$$\lim_{n\to\infty} \left(\int \sum_z u(z) f(z \mid x') d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) f(z \mid x) d\mu_n(u,x) \right) = 0$$

for each $f \in F_c(X)$. Since X is a compact metric space, f is uniformly continuous. Therefore, for any $\eta > 0$, there exists N such that $\max_z |f(z \mid x) - f(z \mid x')| < \eta$ whenever $d(x,x') \leq 1/N$. For every $n \geq N$, we have

$$\left| \int \sum_z u(z) f(z \mid x') d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) f(z \mid x) d\mu_n(u,x) \right|$$
$$\leq \left| \int \sum_z u(z) f(z \mid x') d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) f(z \mid x) d\mu'_n(u,x,x',\omega) \right|$$
$$+ \left| \int \sum_z u(z) f(z \mid x) d\mu'_n(u,x,x',\omega) - \int \sum_z u(z) f(z \mid x) d\mu_n(u,x) \right|.$$

The first term is bounded by $\eta \max_{u,z,z'} |u(z) - u(z')|$; the second term is equal to zero since $\mathrm{mrg}_{1,2}\, \mu'_n = \mu_n$. ■

We can now complete the proof of Lemma 1. Claims 2 and 3 contradict the assumption that $\mathrm{mrg}_{1,3,4}\, \mu'_n$ weakly prefers $O_n(\cdot \mid m'_n, \cdot, \cdot)$ to $O_n(\cdot \mid m_n, \cdot, \cdot)$. ■

Proof of Claim 1. The proof is by induction on n. Suppose that for each $k \leq n-1$, $m_i \in R_i^{n-1}(t_i; \mathcal{T}, \mathcal{M})$ implies $d_{i,k}(\hat{\mu}_{i,k}(t_i), m_{i,k}) \leq \varepsilon_k \leq \varepsilon_{n-1}$ for any agent $i \in I$ and any type $t_i \in T_i$. Suppose that there exists $m_i^* \in R_i^n(t_i; \mathcal{T}, \mathcal{M})$ such that $d_{i,n}(\hat{\mu}_{i,n}(t_i), m_{i,n}^*) > \varepsilon_n$. Then there exists $\nu_i \in \Delta(M_{-i} \times U_i \times T_{-i})$ such that $\nu_i(\{(m_{-i}, u_i, t_{-i}) \mid m_{-i} \in R_{-i}^{n-1}(t_{-i}; \mathcal{T}, \mathcal{M})\}) = 1$, $\mathrm{mrg}_{U_i \times T_{-i}} \nu_i = \mu_i(t_i)$, and ν_i weakly prefers $O(\cdot \mid m_i^*, \cdot)$ to $O(\cdot \mid m'_i, \cdot)$ for any $m'_i \in M_i$.

Collect all the terms in O that depend on $m_{i,n}$, and define $O_{i,n} \colon M_{i,n} \times M_{-i} \to \Delta(Z)$ by

$$
\begin{aligned}
&O_{i,n}(z \mid m_{i,n}, m_{-i}) \\
&= \alpha \left(O_{i,n}^0(z \mid m_{i,n}, m_{-i,1}, \ldots, m_{-i,n-1}) + \sum_{j \in \mathcal{I} \setminus \{i\}} \sum_{k=n+1}^{N} \delta^{k-n} O_{j,k}^0(z \mid m_{j,k}, m_{-j,1}, \ldots, m_{-j,k-1}) \right),
\end{aligned}
$$

where $m_{i,k} = m_{i,k}^*$ for $k \neq n$ when they appear in the second term, and

$$
\alpha = 1 / \left(1 + (|I| - 1) \sum_{k=n+1}^{N} \delta^{k-n} \right)
$$

is a normalization constant. Let $\Omega = \prod_{k=n}^{N} M_{-i,k}$. Since we chose sufficiently small δ, we have $\| O_{i,n}(\cdot \mid \cdot, \cdot, \omega) - O_{i,n}^0 \| \leq \varepsilon_0 \leq \varepsilon_{n-1}$ for any $\omega \in \Omega$. Let $\nu_i^* \in \Delta(M_{-i} \times U_i \times H_{-i,n-1})$ be such that

$$
\nu_i^*(E) = \nu_i(\{(m_{-i}, u_i, t_{-i}) \mid (m_{-i}, u_i, \hat{\mu}_{-i,1}(t_{-i}), \ldots, \hat{\mu}_{-i,n-1}(t_{-i})) \in E\})
$$

for any measurable $E \subseteq M_{-i} \times U_i \times H_{-i,n-1}$. By the induction hypothesis,

$$
\nu_i^*(\{(m_{-i}, u_i, t_{-i,1}, \ldots, t_{-i,n-1}) \mid \max_{j \neq i} \max_{1 \leq k \leq n-1} d_{j,k}(t_{j,k}, m_{j,k}) \leq \varepsilon_{n-1}\}) = 1.
$$

We also have $\mathrm{mrg}_{U_i \times H_{-i,n-1}} \nu_i^* = \hat{\mu}_{i,n}(t_i)$. Thus, we have $\mathrm{mrg}_{M_{-i} \times U_i} \nu_i^* \in \Delta_{\varepsilon_{n-1}, \hat{\mu}_{i,n}(t_i)}(M_{-i} \times U_i)$. Since $M_{i,n}$ is ε_{n-1}-dense in $\Delta(U_i \times H_{-i,n-1})$, there exists $m'_{i,n} \in M_{i,n}$ such that $d_{i,n}(\hat{\mu}_{i,n}(t_i), m'_{i,n}) \leq \varepsilon_{n-1}$. By Lemma 1, $\mathrm{mrg}_{M_{-i} \times U_i} \nu_i^*$ strictly prefers $O_{i,n}(\cdot \mid m'_{i,n}, \cdot)$ to $O_{i,n}(\cdot \mid m_{i,n}^*, \cdot)$, thus $\mathrm{mrg}_{M_{-i} \times U_i} \nu_i^*$ strictly prefers $O(\cdot \mid m'_{i,n}, m_{i,-n}^*, \cdot)$ to $O^*(\cdot \mid m_i^*, \cdot)$. This is a contradiction. ■

B Supplemental Appendix

We present a formal treatment of general interdependent expected utility preferences and the λ-continuity restriction. One way to define state-dependent expected utility preferences for a general measurable space X is to have a preference $\succsim$ over acts over X represented by a belief $\mu \in \Delta(X)$ and a μ-integrable state-dependent utility $u\colon X \times Z \to \mathbb{R}$ as follows:

$$f \succsim f' \Leftrightarrow \int_X \sum_{z \in Z} u(z \mid x)(f(z \mid x) - f'(z \mid x))\mu(dx) \geq 0.$$

Instead, we use a "signed measure" over $X \times Z$, a real-valued countably additive set function, to represent $\succsim$ as

$$f \succsim f' \Leftrightarrow \int_{X \times Z} (f(z \mid x) - f'(z \mid x))\nu(dx, dz) \geq 0.$$

The representation by a signed measure ν is formally equivalent to, via the Radon-Nikodym theorem, but more convenient than the representation by a belief-utility pair (μ, u). For example, u is meaningful only up to μ-null events, and hence multiple belief-utility pairs can represent the same preference. Indeed, although multiple signed measures can also represent the same preference, it is not difficult to pick a particular normalization. For example, if $\succsim$ is not completely indifferent over all outcomes, then we can choose $z, z' \in Z$ such that $z \succ z'$ and represent $\succsim$ uniquely by a signed measure ν over $X \times Z$ such that $\nu(X \times \{z\}) = 1$ and $\nu(E \times \{z'\}) = 0$ for any $E \subseteq X$.

In what follows, we use state-dependent expected utility preferences, briefly described above, to define type spaces of interdependent preferences, preference hierarchies, and the universal type space. Along the way, we introduce various notions directly based on preferences so that we can guarantee easily that these notions are well defined and independent of representations and normalizations. But we also rephrase these notions, whenever possible, in terms of signed-measure representations to ease the reader into possibly unfamiliar notations.

Our exercise here is largely guided by the analogy between subjective beliefs and preferences, originated by Savage (1954) in single-agent environments and extended by Epstein and Wang (1996), di Tillio (2008) and Ganguli and Heifetz (2012) to multi-agent environments. At a technical level, our argument relies on mathematical similarities between probability measures and signed measures. At some subtle level, however, we need to understand a "patchwork" of possibly multiple signed-measure representations of a single preference, which we will discuss further in Section B.5.

B.1 State-Dependent Expected Utility Preferences

For a measurable space X (implicitly endowed with its σ-algebra), let $\text{ca}(X)$ be the set of all finite signed measures over X. For $\nu \in \text{ca}(X)$, $\|\nu\| = \sup_{E,E' \subseteq X}(\nu(E) - \nu(E')) < \infty$ denotes the total

variation of ν; $|\nu|$ denotes the total variation measure on X, defined by $|\nu|(E) = \|\nu(\cdot \cap E)\|$ for each $E \subseteq X$. If X is a compact metric space (implicitly endowed with its Borel σ-algebra), $\mathrm{ca}(X)$ is the dual of the set of continuous functions with the sup norm (the Riesz representation theorem).

Recall that $F(X)$ denotes the set of all acts over X. If X is a compact metric space, $F_c(X) \subseteq F(X)$ denotes the set of all continuous acts over X.

Let $P(X)$ be the set of all state-dependent expected utility preferences over $F(X)$ represented by $\nu \in \mathrm{ca}(X \times Z)$ as follows:

$$f \succsim f' \Leftrightarrow \int_{X \times Z} (f(z \mid x) - f'(z \mid x))\nu(dx, dz) \geq 0.$$

We say that a preference $\succsim \in P(X)$ is *certain of* $E \subseteq X$ if $X \setminus E$ is Savage-null with respect to $\succsim$. For a preference $\succsim \in P(X)$ represented by $\nu \in \mathrm{ca}(X \times Z)$, $\succsim$ is certain of E if and only if $\nu(E' \times \{z\}) = \nu(E' \times \{z'\})$ for any $E' \subseteq X \setminus E$ and $z, z' \in Z$.

We endow $P(X)$ with the σ-algebra generated by $\{\succsim \in P(X) \mid f \succsim f'\}$ for any $f, f' \in F(X)$. If X is a compact metric space, we also endow $P(X)$ with the topology generated by $\{\succsim \in P(X) \mid f \succ f'\}$ for any $f, f' \in F_c(X)$; in this case, the Borel σ-algebra on $P(X)$ coincides with the original σ-algebra on $P(X)$.[27]

Given two measurable spaces X and Y, a measurable mapping $\varphi \colon X \to Y$ and a preference $\succsim \in P(X)$, we can define the *induced preference* $\varphi^P(\succsim)$ as the preference over $F(Y)$ such that for any $f, f' \in F(Y)$, it weakly prefers f to f' if and only if $\succsim$ weakly prefers $f \circ \varphi$ to $f' \circ \varphi$. It is easy to show that if $\succsim \in P(X)$ is represented by a signed measure $\nu \in \mathrm{ca}(X \times Z)$, then the induced preference $\varphi^P(\succsim)$ is represented by the induced signed measure $\nu \circ (\varphi^{-1}, \mathrm{id}_Z) \in \mathrm{ca}(Y \times Z)$. We thus have $\varphi^P(\succsim) \in P(Y)$. Note that $\varphi^P \colon P(X) \to P(Y)$ is measurable; moreover, if X and Y are compact metric spaces and $\varphi \colon X \to Y$ is continuous, then $\varphi^P \colon P(X) \to P(Y)$ is also continuous.

The "marginal" is an important example of induced preferences. Given a product measurable space $X \times Y$ and a preference $\succsim \in P(X \times Y)$, the projection mapping from $X \times Y$ to X induces the *marginal* of $\succsim$, denoted by $\mathrm{mrg}_X \succsim \in P(X)$. In other words, we first identify $F(X)$ as a subset of $F(X \times Y)$, where outcomes do not depend on the Y-coordinate, and then define the marginal of $\succsim$ as the preference over $F(X)$ that can be identified with the restriction of $\succsim$ to such Y-independent acts in $F(X \times Y)$. This notion corresponds to the notion of marginal of a probability or signed measure. Indeed, if $\succsim$ is represented by a signed measure $\nu \in \mathrm{ca}(X \times Y \times Z)$, then $\mathrm{mrg}_X \succsim$ is represented by

[27] Since $F_c(X) \subseteq F(X)$, any Borel-measurable subset of $P(X)$ is measurable. Conversely, let $\mathcal{D} = \{E \subseteq X \mid \{\succsim \in P(X) \mid y_E y' \succsim y''_E y'''\}$ is Borel-measurable for any $y, y', y'', y''' \in \Delta(Z)\}$, where $y_E y'$ denotes the act that takes values y on E and y' on $X \setminus E$. Then $\mathcal{D}$ is a Dynkin system, and contains all closed subsets of X by Urysohn's lemma. Thus $\mathcal{D}$ coincides with the Borel σ-algebra on X, and hence $\{\succsim \in P(X) \mid f \succsim f'\}$ is Borel-measurable if f and f' are in the form of $y_E y'$ with Borel-measurable $E \subseteq X$. This extends to all simple acts and to all acts in the usual way.

the marginal of ν on $X \times Z$, $\mathrm{mrg}_{X \times Z} \nu \in \mathrm{ca}(X \times Z)$, where $(\mathrm{mrg}_{X \times Z} \nu)(E \times \{z\}) = \nu(E \times Y \times \{z\})$ for any $E \subseteq X$ and $z \in Z$.

For a more specific example, consider a measurable space X, an arbitrary singleton set $\{*\}$ and a preference $\succsim \in P(X)$. Then the constant mapping from X to $\{*\}$ induces the unconditional preference of $\succsim$ over lotteries. If $\nu \in \mathrm{ca}(X \times Z)$ represents $\succsim$, then $\mathrm{mrg}_Z \nu \in \mathrm{ca}(Z) \cong \mathbb{R}^Z$ is a von Neumann-Morgenstern utility index that represents the unconditional preference of $\succsim$ over lotteries.

B.2 Type Spaces and the Universal Type Space

A *type space* is given by $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$, where, for each $i \in \mathcal{I}$, T_i is a measurable space of agent i's types, and $\pi_i \colon T_i \to P(T_{-i})$ is a measurable mapping that maps his types to preferences.

Let $H_0 = \{*\}$ (an arbitrary singleton set) and $H_n = H_{n-1} \times P(H_{n-1}^{|I|-1}) = \prod_{k=0}^{n-1} P(H_k^{|I|-1})$ for each $n \geq 1$. Let $H = \prod_{n=0}^{\infty} P(H_n^{|I|-1})$ be the set of all hierarchies of preferences.

Given a type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$, we define the *preference hierarchy* of a type $t_i \in T_i$, $\hat{\pi}_i(t_i) = (\hat{\pi}_{i,1}(t_i), \hat{\pi}_{i,2}(t_i), \ldots)$, as follows: $\hat{\pi}_{i,1}(t_i)$ is the unconditional preference of $\pi_i(t_i)$ over lotteries, and for each $n \geq 2$, $\hat{\pi}_{i,n}(t_i)$ is the preference of type t_i over acts over the opponents' first $(n-1)$ order preferences, i.e., $\hat{\pi}_{i,n}(t_i) = (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1})^P(\pi_i(t_i))$. It is easy to show inductively that $\hat{\pi}_{i,n} \colon T_i \to P(H_{n-1}^{|I|-1})$ is measurable for any $n \geq 1$, and hence $\hat{\pi}_i \colon T_i \to \prod_{n=0}^{\infty} P(H_n^{|I|-1})$ is also measurable.

Following Heifetz and Samet (1998), we define T_i^* as the set of all preference hierarchies $h \in H$ such that $h = \hat{\pi}_i(t_i)$ for some type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ and some type $t_i \in T_i$. We define $\pi_i^* \colon T_i^* \to P(T_{-i}^*)$ by

$$\pi_i^*(t_i^*) = \hat{\pi}_{-i}^P(\pi_i(t_i)).$$

We can show that π_i^* is well defined (i.e., independent of particular type space $\mathcal{T}$ and particular type t_i) and measurable.[28] We thus have the *universal type space* $\mathcal{T}^* = (T_i^*, \pi_i^*)_{i \in I}$. By construction, a

[28] For each $n \geq 0$, let $\mathrm{pr}_{-i,n} \colon T_i^* \to H_n^{|I|-1}$ be the projection mapping. Fix any $f, f' \in F(H_n^{|I|-1})$. For each $t_i^* = (\succsim_1, \succsim_2, \ldots) \in T_i^*$, there exist a type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ and a type $t_i \in T_i$ such that $t_i^* = \hat{\pi}_i(t_i)$. Then we have

$$\begin{aligned}
&\pi_i^*(t_i^*) \text{ weakly prefers } f \circ \mathrm{pr}_{-i,n} \text{ to } f' \circ \mathrm{pr}_{-i,n} \\
&\Leftrightarrow \pi_i(t_i) \text{ weakly prefers } f \circ \hat{\pi}_{-i,n}^P \text{ to } f' \circ \hat{\pi}_{-i,n}^P \\
&\Leftrightarrow \succsim_{n+1} \text{ weakly prefers } f \text{ to } f'.
\end{aligned}$$

Thus $\{t_i^* \in T_i^* \mid \pi_i^*(t_i^*) \text{ weakly prefers } f \circ \mathrm{pr}_{-i,n} \text{ to } f' \circ \mathrm{pr}_{-i,n}\}$ is well defined and measurable. Since this is true for any n and any $f, f' \in F(H_n^{|I|-1})$, $\{t_i^* \in T_i^* \mid \pi_i^*(t_i^*) \text{ weakly prefers } f \text{ to } f'\}$ is well defined and measurable for any $f, f' \in F(T_{-i}^*)$, and hence $\pi_i^* \colon T_i^* \to P(T_{-i}^*)$ is well defined and measurable.

profile $(\hat{\pi}_i)_{i\in I}$ of mappings is a preference-preserving morphism, also known as a type morphism in Heifetz and Samet (1998), from any type space $\mathcal{T}$ to $\mathcal{T}^*$ in the following sense.[29]

Proposition 7 *For any type space $\mathcal{T} = (T_i, \pi_i)_{i\in I}$ and any $i \in I$, we have $\pi_i^* \circ \hat{\pi}_i = \hat{\pi}_{-i}^P \circ \pi_i$.*

B.3 Proof of Proposition 4

The following result drops the simplex assumption in Proposition 3.

Proposition 8 *For any type space $\mathcal{T} = (T_i, \pi_i)_{i\in I}$, any agent $i \in I$, and any type $t_i \in T_i$, we have*

$$E_i(t_i; \mathcal{T}, \mathcal{M}) \supseteq E_i(\hat{\pi}_i(t_i); \mathcal{T}^*, \mathcal{M})$$

for any mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$.

Proof. The proof is analogous to that of Proposition 3; we only need to replace the belief-preserving property by the preference-preserving property established in Proposition 7. ■

Proposition 4 follows from Proposition 8 and the existence of equilibria for countable types.

B.4 Proof of Proposition 5

Given a type space $\mathcal{T} = (T_i, \pi_i)_{i\in I}$ and a mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$, we define the set of rationalizable actions, denoted by $R_i(t_i)$ or $R_i(t_i; \mathcal{T}, \mathcal{M})$ as follows:

$$R_i^0(t_i) = M_i,$$

$$R_i^{n+1}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \succsim_i \in P(M_{-i} \times T_{-i}) \text{ s.t.} \\ \text{(i) } \succsim_i \text{ is certain of the graph of } R_{-i}^n, \\ \text{(ii) } \operatorname{mrg}_{T_{-i}} \succsim_i = \pi_i(t_i), \\ \text{(iii) } \succsim_i \text{ weakly prefers } O(\cdot \mid m_i, \cdot) \text{ to } O(\cdot \mid m_i', \cdot), \quad \text{for any } m_i' \in M_i. \end{array} \right\},$$

$$R_i(t_i) = \bigcap_{n=0}^{\infty} R_i^n(t_i).$$

[29] In passing, we note that any preference-preserving morphism preserves preference hierarchies, and that $(\hat{\pi}_i)_{i\in I}$ is the unique preference-preserving morphism from $\mathcal{T}$ to $\mathcal{T}^*$.

Lemma 2 *For any finite type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and any mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$, we have the following:*

1. *We have $m_i \notin R_i^1(t_i)$ if and only if there exists $\sigma_i \in \Delta(M_i)$ such that*
   (a) *$O(\cdot \mid \sigma_i, m_{-i}) - O(\cdot \mid m_i, m_{-i})$ is independent of $m_{-i} \in M_{-i}$, and*
   (b) *$\pi_i(t_i)$ strictly prefers $O(\cdot \mid \sigma_i, m_{-i})$ to $O(\cdot \mid m_i, m_{-i})$ for some (and hence for all) $m_{-i} \in M_{-i}$.*[30]
2. *$R_i(t_i) = R_i^1(t_i)$.*

Proof. For part 1, the if direction is immediate. To show the only-if direction, let $\pi_i(t_i)$ be represented by $\bar{w}_i \colon T_{-i} \times Z \to \mathbb{R}$ as follows:

$$f \succsim f' \Leftrightarrow \sum_{t_{-i}, z} (f(z \mid t_{-i}) - f'(z \mid t_{-i})) \bar{w}_i(t_{-i}, z) \geq 0.$$

If $m_i \notin R_i^1(t_i)$, then there is no $w_i \colon M_{-i} \times T_{-i} \times Z \to \mathbb{R}$ such that

$$\sum_{m_{-i}} w_i(m_{-i}, t_{-i}, z) = \bar{w}_i(t_{-i}, z) \qquad \text{for all } t_{-i}, z,$$

$$\sum_{m_{-i}, t_{-i}, z} (O(z \mid m_i, m_{-i}) - O(z \mid m_i', m_{-i})) w_i(m_{-i}, t_{-i}, z) \geq 0 \qquad \text{for all } m_i'.$$

By Farkas' lemma, there exist $D \colon T_{-i} \times Z \to \mathbb{R}$ and $\sigma_i \in \Delta(M_i)$ such that

$$D(z \mid t_{-i}) - (O(z \mid \sigma_i, m_{-i}) - O(z \mid m_i, m_{-i})) = 0 \qquad \text{for all } t_{-i}, m_{-i}, z,$$

$$\sum_{t_{-i}, z} D(z \mid t_{-i}) \bar{w}_i(t_{-i}, z) > 0.$$

Thus $O(\cdot \mid \sigma_i, m_{-i}) - O(\cdot \mid m_i, m_{-i})$ is independent of m_{-i}, and $\pi_i(t_i)$ strictly prefers $O(\cdot \mid \sigma_i, m_{-i})$ to $O(\cdot \mid m_i, m_{-i})$.

For part 2, fix any player $i \in I$. For each $j \neq i$ and $t_j \in T_j$, if $m_j \in R_j^1(t_j)$, then let $\sigma_j(\cdot \mid m_j, t_j)$ be the point mass on m_j. If $m_j \notin R_j^1(t_j)$, then by part 1, there exists $\sigma_j(\cdot \mid m_j, t_j) \in \Delta(M_j)$ such that for any $z \in Z$, $O(z \mid \sigma_j(\cdot \mid m_j, t_j), m_{-j}) - O(z \mid m_j, m_{-j})$ is independent of m_{-j}. Without loss of generality, we assume that $\sigma_j(\cdot \mid m_j, t_j) \in \Delta(R_j^1(t_j))$. For each $m_{-i} \in M_{-i}$ and $t_{-i} \in T_{-i}$,

[30] We define $O(\cdot \mid \sigma_i, m_{-i})$ by

$$O(z \mid \sigma_i, m_{-i}) = \sum_{m_i'} O(z \mid m_i', m_{-i}) \sigma_i(m_i')$$

for each $z \in Z$.

define $\sigma_{-i}(\cdot \mid m_{-i}, t_{-i}) \in \Delta(R^1_{-i}(t_{-i}))$ by $\sigma_{-i}(m'_{-i} \mid m_{-i}, t_{-i}) = \prod_{j \neq i} \sigma_j(m'_j \mid m_j, t_j)$ for each $m'_{-i} \in R^1_{-i}(t_{-i})$.

Pick any $t_i \in T_i$ and any $m_i \in R^1_i(t_i)$. Then there exists $\succsim_i \in P(M_{-i} \times T_{-i})$ such that $\mathrm{mrg}_{T_{-i}} \succsim_i = \pi_i(t_i)$ and m_i is a best response with respect to $\succsim_i$. We will show that m_i survives in the second step of iteration.

Let $\pi_i(t_i)$ be represented by $\bar{w}_i \colon T_{-i} \times Z \to \mathbb{R}$. Let $\succsim_i$ be represented by $w_i \colon M_{-i} \times T_{-i} \times Z \to \mathbb{R}$ such that $\sum_{m_{-i}} w_i(m_{-i}, \cdot, \cdot) = \bar{w}_i$. Define $w'_i \colon M_{-i} \times T_{-i} \times Z \to \mathbb{R}$ by

$$w'_i(m'_{-i}, t_{-i}, z) = \sum_{m_{-i}} \sigma_{-i}(m'_{-i} \mid m_{-i}, t_{-i}) w_i(m_{-i}, t_{-i}, z)$$

for $m'_{-i} \in M_{-i}$, $t_{-i} \in T_{-i}$ and $z \in Z$. Denote by $\succsim'_i \in P(M_{-i} \times T_{-i})$ the preference represented by w'_i. First, since $\sigma_{-i}(\cdot \mid m_{-i}, t_{-i}) \in \Delta(R^1_{-i}(t_{-i}))$ for any $m_{-i} \in M_{-i}$ and $t_{-i} \in T_{-i}$, $\succsim'_i$ is certain of the graph of R^1_{-i}. Second, we have

$$\begin{aligned}
\sum_{m'_{-i}} w'_i(m'_{-i}, t_{-i}, z) &= \sum_{m_{-i}, m'_{-i}} \sigma_{-i}(m'_{-i} \mid m_{-i}, t_{-i}) w_i(m_{-i}, t_{-i}, z) \\
&= \sum_{m_{-i}} w_i(m_{-i}, t_{-i}, z) \\
&= \bar{w}_i(t_{-i}, z),
\end{aligned}$$

and hence $\mathrm{mrg}_{T_{-i}} \succsim'_i = \pi_i(t_i)$. Third, for any $m'_i \in M_i$, we have

$$\begin{aligned}
&\sum_{m'_{-i}} O(z \mid m'_i, m'_{-i}) w'_i(m'_{-i}, t_{-i}, z) \\
&= \sum_{m_{-i}, m'_{-i}} O(z \mid m'_i, m'_{-i}) \sigma_{-i}(m'_{-i} \mid m_{-i}, t_{-i}) w_i(m_{-i}, t_{-i}, z) \\
&= \sum_{m_{-i}} O(z \mid m'_i, \sigma_{-i}(\cdot \mid m_{-i}, t_{-i})) w_i(m_{-i}, t_{-i}, z) \\
&= \sum_{m_{-i}} (O(z \mid m'_i, m_{-i}) + D(z \mid m_{-i}, t_{-i})) w_i(m_{-i}, t_{-i}, z) \\
&= \sum_{m_{-i}} O(z \mid m'_i, m_{-i}) w_i(m_{-i}, t_{-i}, z) + \sum_{m_{-i}} D(z \mid m_{-i}, t_{-i}) w_i(m_{-i}, t_{-i}, z)
\end{aligned}$$

for any $t_{-i} \in T_{-i}$ and $z \in Z$, where $D(z \mid m_{-i}, t_{-i}) := O(z \mid m'_i, \sigma_{-i}(\cdot \mid m_{-i}, t_{-i})) - O(z \mid m'_i, m_{-i})$ is independent of m'_i by the definition of $\sigma_{-i}(\cdot \mid m_{-i}, t_{-i})$. Since m_i is a best response with respect to $\succsim_i$ represented by w_i, it is also a best response with respect to $\succsim'_i$ represented by w'_i. ■

Lemma 3 *For any two finite type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T'_i, \pi'_i)_{i \in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t'_i \in T'_i$, if $\hat{\pi}_{i,1}(t_i; \mathcal{T}) = \hat{\pi}_{i,1}(t'_i; \mathcal{T}')$, then we have*

$$R_i(t_i; \mathcal{T}, \mathcal{M}) = R_i(t'_i; \mathcal{T}', \mathcal{M})$$

for any mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$.

Proof. Follows from Lemma 2. ■

Proposition 5 follows from rewriting the statement of Lemma 3 in terms of equilibrium.

B.5 Compactness and Metrizability of $P_\lambda(X)$

Let $P_0(X)$ be the set of preferences in $P(X)$ that are not completely indifferent over all outcomes. By excluding the preference that is completely indifferent over $F(X)$, we can show that $P_0(X)$ is Hausdorff if X is a compact metric space.[31]

Lemma 4 *If X is a compact metric space, then $P_0(X)$ is Hausdorff.*

Proof. Pick any pair of preferences $\succsim, \succsim' \in P_0(X)$ such that $\succsim \neq \succsim'$. Then there exist $f, f' \in F(X)$ such that $\succsim$ and $\succsim'$ have different preferences between f and f'. Since neither $\succsim$ nor $\succsim'$ is completely indifferent, we can assume without loss of generality that $f \succ f'$ and $f' \succ' f$.[32]

Let $\nu, \nu' \in \mathrm{ca}(X \times Z)$ be finite signed measures that represent $\succsim$ and $\succsim'$, respectively. Applying Lusin's theorem to $(X \times Z, |\nu| + |\nu'|)$, we can assume without loss of generality that $f, f' \in F_c(X)$. Thus $\succsim$ and $\succsim'$ are separated by two disjoint open sets generated by f and f'. ■

We define λ-continuity as follows.

Definition 1 *We say that a preference $\succsim$ is λ-continuous with $\lambda > 0$ if there exist $z, z' \in Z$ such that $z \succ z'$ and $(1-\lambda) z + \lambda f \succsim (1-\lambda) z' + \lambda f'$ for any $f, f' \in F(X)$.*

If X is a compact metric space, then by Lusin's theorem, we can require $(1-\lambda) z + \lambda f \succsim (1-\lambda) z' + \lambda f'$ only for any $f, f' \in F_c(X)$ without loss of generality.

Let $P_{z,z',\lambda}(X)$ be the set of all λ-continuous preferences for a fixed pair of outcomes $z, z' \in Z$. Let $P_\lambda(X) = \bigcup_{z,z' \in Z} P_{z,z',\lambda}(X)$ be the set of all λ-continuous preferences.

Note that λ-continuity is preserved for induced preferences. That is, given any measurable mapping $\varphi \colon X \to Y$, if we have $\succsim \in P_{z,z',\lambda}(X)$, then we also have $\varphi^P(\succsim) \in P_{z,z',\lambda}(Y)$; if we have $\succsim \in P_\lambda(X)$, then we also have $\varphi^P(\succsim) \in P_\lambda(Y)$.

[31] Indeed, Lemma 4 holds as long as we exclude the preference that is completely indifferent over $F(X)$, but excluding all preferences that are completely indifferent over all outcomes is crucial for establishing Lemma 7 and hence Proposition 10.

[32] For example, if $f \sim f'$ and $f' \succ' f$, then pick $f'', f''' \in F(X)$ such that $f'' \succ f'''$. Then by slightly mixing f with f'' and f' with f''', we can make the first preference relation strict while maintaining the second preference relation.

Fix a pair of outcomes $z, z' \in Z$ and $\lambda > 0$. Then each $\succsim \in P_{z,z',\lambda}(X)$ is uniquely represented by $\nu \in \mathrm{ca}_{z,z',\lambda}(X \times Z)$, where

$$\mathrm{ca}_{z,z',\lambda}(X \times Z) = \left\{ \nu \in \mathrm{ca}(X \times Z) \;\middle|\; \begin{array}{l} \nu(X \times \{z\}) = 1 \\ \nu(E \times \{z'\}) = 0 \text{ for any } E \subseteq X \\ \int_{X \times Z} (f(z \mid x) - f'(z \mid x)) \nu(dx, dz) \leq (1 - \lambda)/\lambda \\ \text{ for any } f, f' \in F(X) \end{array} \right\}.$$

In words, we normalize a signed-measure representation by first shifting the ex post utility of getting z' conditional on any event to 0, and then scaling the expected utility of getting z to 1. The condition that $\int_{X \times Z} (f(z \mid x) - f'(z \mid x)) \nu(dx, dz) \leq (1 - \lambda)/\lambda$ for any $f, f' \in F(X)$ is a rewriting of the definition of λ-continuity in terms of signed-measure representations. Via this normalization, $P_{z,z',\lambda}(X)$ is measurably isomorphic to $\mathrm{ca}_{z,z',\lambda}(X \times Z)$; furthermore, if X is a compact metric space, then $P_{z,z',\lambda}(X)$ is topologically isomorphic (i.e., homeomorphic) to $\mathrm{ca}_{z,z',\lambda}(X \times Z)$ endowed with the weak* topology.

Note that this normalization is preserved for induced preferences. That is, given any measurable mapping $\varphi \colon X \to Y$, if ν belongs to $\mathrm{ca}_{z,z',\lambda}(X \times Z)$, then the induced signed measure $\nu \circ (\varphi^{-1}, \mathrm{id}_Z)$ belongs to $\mathrm{ca}_{z,z',\lambda}(Y \times Z)$ with the same $z, z' \in Z$ and $\lambda > 0$. Therefore, if $\succsim \in P_{z,z',\lambda}(X)$ is represented by a normalized signed measure $\nu \in \mathrm{ca}_{z,z',\lambda}(X \times Z)$, then the induced preference $\varphi^P(\succsim) \in P_{z,z',\lambda}(Y)$ is represented by the already normalized signed measure $\nu \circ (\varphi^{-1}, \mathrm{id}_Z) \in \mathrm{ca}_{z,z',\lambda}(Y \times Z)$.

Since any measurable function $g \colon X \times (Z \setminus \{z'\}) \to \mathbb{R}$ with $\|g\| \leq 1/(|Z| - 1)$ in the sup norm can be written as $g(x, z) = f(z \mid x) - f'(z \mid x)$ with some $f, f' \in F(X)$, we have $\|\nu\| \leq (|Z| - 1)(1 - \lambda)/\lambda$ for any $\nu \in \mathrm{ca}_{z,z',\lambda}(X \times Z)$. Conversely, since $\|f - f'\| \leq 1$ in the sup norm for any $f, f' \in F(X)$, we have $\nu \in \mathrm{ca}_{z,z',\lambda}(X \times Z)$ for any $\nu \in \mathrm{ca}(X \times Z)$ such that $\nu(X \times \{z\}) = 1$, $\nu(E \times \{z'\}) = 0$ for any $E \subseteq X$, and $\|\nu\| \leq (1 - \lambda)/\lambda$.

Lemma 5 *If X is a compact metric space, then $P_{z,z',\lambda}(X)$ is compact and metrizable for any $z, z' \in Z$ and $\lambda > 0$.*

Proof. By the remark after Definition 1, $P_{z,z',\lambda}(X)$ is closed in $P_0(X)$. Also, $\mathrm{ca}_{z,z',\lambda}(X \times Z)$ can be seen as a subset of the ball $\{\nu \in \mathrm{ca}(X \times (Z \setminus \{z'\})) \mid \|\nu\| \leq (|Z| - 1)(1 - \lambda)/\lambda\}$, which is weak*-compact by the Riesz representation theorem and Alaoglu's theorem, and weak*-metrizable by the Stone-Weierstrass theorem. Thus $\mathrm{ca}_{z,z',\lambda}(X \times Z)$ is compact and metrizable, and so is $P_{z,z',\lambda}(X)$. ■

Note that Lemma 5 relies on λ-continuity with $\lambda > 0$.

Recall that $P_\lambda(X) = \bigcup_{z,z' \in Z} P_{z,z',\lambda}(X)$. If $|Z| \geq 3$, then this union is not disjoint, i.e., a given preference $\succsim \in P_\lambda(X)$ may belong to $P_{z,z',\lambda}(X)$ with multiple pairs of (z, z'). In this case, we do not choose any specific (z, z') pair as "canonical". Instead, we view $P_\lambda(X)$ as a "patchwork" of finitely many $P_{z,z',\lambda}(X)$, each of which is homeomorphic to $\mathrm{ca}_{z,z',\lambda}(X \times Z)$.

Proposition 9 *If X is a compact metric space, then $P_\lambda(X)$ is compact and metrizable for any $\lambda > 0$.*

Proof. By Lemmas 4 and 5, $P_\lambda(X)$ is a finite union of compact and metrizable subspaces $P_{z,z',\lambda}(X)$, and hence $P_\lambda(X)$ is compact and metrizable. (The metrizability follows from the Nagata-Smirnov metrization theorem. See Nagata (1985, Theorem 6.12).) ■

B.6 The Robust Scoring Rule

As in Section 3.5.1, we analyze a single-agent mechanism that reveals her state-dependent preferences. Fix $\lambda > 0$. Fix a compact metric space X with metric d. By Proposition 9, $P_\lambda(X)$ is also a compact metric space, whose metric is denoted by d_P. The choice function with respect to $\succsim \in P_\lambda(X)$ is given by

$$C_{\succsim}(f, f') = \begin{cases} f & \text{if } \succsim \text{ weakly prefers } f \text{ to } f', \\ f' & \text{if } \succsim \text{ strictly prefers } f' \text{ to } f \end{cases}$$

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

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.

We consider the following direct mechanism $\mathcal{M}^0 = (M^0, O^0)$ for a single agent with message set $M^0 = P_\lambda(X)$ and outcome function $O^0 \colon M^0 \times X \to \Delta(Z)$ given by

$$O^0(z \mid m, x) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} 2^{-k-l} C_m(f_k, f_l)(z \mid x) \tag{3}$$

for each realized state $x \in X$ and reported preference $m \in M^0$.

For each $\delta > 0$, $\succsim \in P_\lambda(X)$, and measurable space Ω, let

$$P_{\lambda,\delta,\succsim}(X \times \Omega) = \left\{ \mathrm{mrg}_{2,3} \succsim' \in P_\lambda(X \times \Omega) \mid \begin{array}{l} \exists \succsim' \in P_\lambda(X \times X \times \Omega) \text{ s.t.} \\ \text{(i) } \succsim' \text{ is certain of } \{(x, x', \omega) \mid d(x, x') \leq \delta\}, \\ \text{(ii) } \mathrm{mrg}_1 \succsim' = \succsim \end{array} \right\},$$

where $\mathrm{mrg}_\Lambda \succsim'$ with $\Lambda \subset \{1, 2, 3\}$ denotes the marginal of $\succsim'$ with respect to the coordinates in Λ.

Lemma 6 *Fix $\lambda > 0$. For every $\varepsilon > 0$, there exists $\delta > 0$ such that the following is true for any preference $\succsim \in P_\lambda(X)$, any pair of messages m, m', any measurable space Ω, and any perturbed outcome function $O\colon M^0 \times X \times \Omega \to \Delta(Z)$: if $d_P(\succsim, m) \leq \delta$, $d_\Delta(\succsim, m') > \varepsilon$, and $\|O(\cdot \mid \cdot, \cdot, \omega) - O^0\| \leq \delta$ for any $\omega \in \Omega$, then any preference in $P_{\lambda,\delta,\succsim}(X \times \Omega)$ strictly prefers $O(\cdot \mid m, \cdot, \cdot)$ to $O(\cdot \mid m', \cdot, \cdot)$.*

Proof. Suppose not. Then there exists $\varepsilon > 0$ such that for every $n \in \mathbb{N}$, there exist $\succsim_n, m_n, m'_n \in P_\lambda(X)$ with $d_P(\succsim_n, m_n) \leq 1/n$ and $d_P(\succsim_n, m'_n) > \varepsilon$, measurable space Ω_n, perturbed outcome function $O_n\colon M^0 \times X \times \Omega_n \to \Delta(Z)$ with $\|O_n(\cdot \mid \cdot, \cdot, \omega) - O^0\| \leq 1/n$ for every $\omega \in \Omega_n$, $\succsim'_n \in P_\lambda(X \times X \times \Omega_n)$ such that $\succsim'_n$ is certain of $\{(x, x', \omega) \mid d(x, x') \leq \delta\}$, $\operatorname{mrg}_1 \succsim'_n = \succsim_n$, and $\operatorname{mrg}_{2,3} \succsim'_n$ weakly prefers $O_n(\cdot \mid m'_n, \cdot, \cdot)$ to $O_n(\cdot \mid m_n, \cdot, \cdot)$. By taking a subsequence if necessary, we can assume without loss of generality that $\succsim'_n \in P_{z,z',\lambda}(X \times X \times \Omega_n)$ with a fixed (z, z') pair, and hence $\succsim_n \in P_{z,z',\lambda}(X)$ with the same (z, z') pair. By Proposition 9, by taking a subsequence if necessary, we can find $\succsim^*, m'^* \in P_\lambda(X)$ such that $\succsim_n \to \succsim^*$ and $m'_n \to m'^*$ as $n \to \infty$. Note that $m_n \to \succsim^*$ as $n \to \infty$, and $\succsim^* \neq m'^*$. Also note that $\succsim^* \in P_{z,z',\lambda}(X)$. Let $\nu_n, \nu^* \in \operatorname{ca}_{z,z',\lambda}(X \times Z)$ and $\nu'_n \in \operatorname{ca}_{z,z',\lambda}(X \times X \times \Omega_n \times Z)$ represent $\succsim_n$, $\succsim^*$, and $\succsim'_n$, respectively. Note that $\operatorname{mrg}_{1,4} \nu'_n = \nu_n$.

Let

$$u^* = \int O^0(z \mid \succsim^*, x) d\nu^*(x, z).$$

Claim 4 *We have*

$$\lim_{n\to\infty} \int O^0(z \mid m_n, x) d\nu_n(x, z) = u^*,$$
$$\limsup_{n\to\infty} \int O^0(z \mid m'_n, x) d\nu_n(x, z) < u^*.$$

Proof of Claim 4. The claim follows from showing that

$$\lim_{n\to\infty} \int C_{m_n}(f_k, f_l)(z \mid x) d\nu_n(x, z) = \int C_{\succsim^*}(f_k, f_l)(z \mid x) d\nu^*(x, z),$$
$$\limsup_{n\to\infty} \int C_{m'_n}(f_k, f_l)(z \mid x) d\nu_n(x, z) \leq \int C_{\succsim^*}(f_k, f_l)(z \mid x) d\nu^*(x, z),$$

for each k, l, and that the second inequality holds with strict inequality for some k, l. The first equality and the second weak inequality follow from the standard revealed preference argument. To show the strict inequality, since $\succsim^* \neq m'^*$ and $F \subset F_c(X)$ is dense in the sup norm, there exist k, l such that $\succsim^*$ strictly prefers f_k to f_l while m'^* strictly prefers f_l to f_k. Since m'_n strictly prefers

f_l to f_k for sufficiently large n, we have

$$\begin{aligned}
&\lim_{n\to\infty}\int C_{m'_n}(f_k,f_l)(z\mid x)d\nu_n(x,z)\\
&=\lim_{n\to\infty}\int f_l(z\mid x)d\nu_n(x,z)\\
&=\int f_l(z\mid x)d\nu^*(x,z)\\
&<\int f_k(z\mid x)d\nu^*(x,z)\\
&=\int C_{\gtrsim^*}(f_k,f_l)(z\mid x)d\nu^*(x,z).
\end{aligned}$$

■

Claim 5 *We have*

$$\lim_{n\to\infty}\left(\int O_n(z\mid m,x',\omega)d\nu'_n(x,x',\omega,z)-\int O^0(z\mid m,x)d\nu_n(x,z)\right)=0$$

and the convergence is uniform in $m\in M^0$.

Proof of Claim 5. Note that

$$\begin{aligned}
&\left|\int O_n(z\mid m,x',\omega)d\nu'_n(x,x',\omega,z)-\int O^0(z\mid m,x)d\nu_n(x,z)\right|\\
&\le\left|\int O_n(z\mid m,x',\omega)d\nu'_n(x,x',\omega,z)-\int O^0(z\mid m,x')d\nu'_n(x,x',\omega,z)\right|\\
&\quad+\left|\int O^0(z\mid m,x')d\nu'_n(x,x',\omega,z)-\int O^0(z\mid m,x)d\nu_n(x,z)\right|.
\end{aligned}$$

The first term is bounded by $\sup_{\omega\in\Omega_n}\|O_n(\cdot\mid\cdot,\cdot,\omega)-O^0\|\|\nu'_n\|\le(|Z|-1)(1-\lambda)/(n\lambda)$.

To show that the second term converges to 0 uniformly in m, it is enough to show that

$$\lim_{n\to\infty}\left(\int f(z\mid x')d\nu'_n(x,x',\omega,z)-\int f(z\mid x)d\nu_n(x,z)\right)=0$$

for each $f\in F_c(X)$. Since X is a compact metric space, f is uniformly continuous. Therefore, for any $\eta>0$, there exists N such that $\max_z|f(z\mid x)-f(z\mid x')|<\eta$ whenever $d(x,x')\le 1/N$. For every $n\ge N$, we have

$$\begin{aligned}
&\left|\int f(z\mid x')d\nu'_n(x,x',\omega,z)-\int f(z\mid x)d\nu_n(x,z)\right|\\
&\le\left|\int f(z\mid x')d\nu'_n(x,x',\omega,z)-\int f(z\mid x)d\nu'_n(x,x',\omega,z)\right|\\
&\quad+\left|\int f(z\mid x)d\nu'_n(x,x',\omega,z)-\int f(z\mid x)d\nu_n(x,z)\right|.
\end{aligned}$$

The first term is bounded by $\eta\|\nu'_n\| \leq \eta(|Z|-1)(1-\lambda)/\lambda$; the second term is equal to zero since $\mathrm{mrg}_{1,4}\,\nu'_n = \nu_n$. ■

Claims 4 and 5 contradict the assumption that $\mathrm{mrg}_{2,3}\succsim'_n$ weakly prefers $O_n(\cdot \mid m'_n, \cdot, \cdot)$ to $O_n(\cdot \mid m_n, \cdot, \cdot)$. ■

B.7 λ-Continuous Type Spaces and the Universal λ-Continuous Type Space

We say that a type space $\mathcal{T} = (T_i, \pi_i)_{i\in\mathcal{I}}$ is *λ-continuous* if $\pi_i(t_i) \in P_\lambda(T_{-i})$ for any $i \in I$ and $t_i \in T_i$. As we argued in Section 5, λ-continuity is a weak requirement. A type space is λ-continuous with some $\lambda > 0$ if it satisfies a simplex restriction, or if the type space is finite (Abreu and Matsushima (1992b)). The second sufficient condition can generalize to compact and continuous type spaces as follows. We say that a preference type space $\mathcal{T} = (T_i, \pi_i)_{i\in\mathcal{I}}$ is *compact and continuous* if, for each $i \in \mathcal{I}$, T_i is a compact metric space, and $\pi_i \colon T_i \to P(T_{-i})$ is continuous.

Proposition 10 *If a type space $\mathcal{T} = (T_i, \pi_i)_{i\in\mathcal{I}}$ with $\pi_i \colon T_i \to P_0(T_{-i})$ is compact and continuous, then it is $\lambda_{\mathcal{T}}$-continuous with some $\lambda_{\mathcal{T}} > 0$.*

This follows immediately from the following lemma.

Lemma 7 *Let X be a compact metric space. For any compact subset Q of $P_0(X)$, there exists $\lambda_Q > 0$ such that any preference in Q is λ_Q-continuous.*

Proof. Fix any $\succsim \in Q \subseteq P_0(X)$. Then there exist $z, z' \in Z$ such that $z \succ z'$. Let $\nu \in \mathrm{ca}(X \times Z)$ be a finite signed measure that represents $\succsim$ and is normalized by $\nu(X \times \{z\}) = 1$ and $\nu(E \times \{z'\}) = 0$ for any $E \subseteq X$. Let $\lambda = 1/(\|\nu\|+1) > 0$. Then we have $\|\nu\| = (1-\lambda)/\lambda$, and hence $\nu \in \mathrm{ca}_{z,z',\lambda}(X \times Z)$, i.e., $\succsim \in P_{z,z',\lambda}(X)$. Since a neighborhood of $\succsim$ is contained in $P_{z,z',\lambda/2}(X)$, by the usual compactness argument, we can take desired $\lambda_Q > 0$ uniformly over Q. ■

Let $H_{\lambda,0} = \{*\}$, $H_{\lambda,n} = H_{\lambda,n-1} \times P_\lambda(H_{\lambda,n-1}^{|I|-1})$ for each $n \geq 1$, and $H_\lambda = \prod_{n=0}^\infty P_\lambda(H_{\lambda,n}^{|I|-1})$. By Proposition 9, $H_{\lambda,n}$ is compact and metrizable for any $n \geq 0$. We endow H_λ with the product topology, and hence H_λ is also compact and metrizable. Since λ-continuity is preserved for induced preferences, the preference hierarchy of any λ-continuous type is also λ-continuous. That is, for any λ-continuous type space $\mathcal{T} = (T_i, \pi_i)_{i\in\mathcal{I}}$ and any type $t_i \in T_i$, we have $\hat{\pi}_i(t_i) \in H_\lambda$. (Recall that $\hat{\pi}_i(t_i)$ denotes the preference hierarchy of t_i.) Following Heifetz and Samet (1998), we define $T^*_{i,\lambda}$ as the set of all preference hierarchies $h \in H_\lambda$ such that $h = \hat{\pi}_i(t_i)$ for some λ-continuous type space $\mathcal{T} = (T_i, \pi_i)_{i\in\mathcal{I}}$ and some type $t_i \in T_i$. Then we have the *universal λ-continuous type space* $\mathcal{T}^*_\lambda = (T^*_{i,\lambda}, \pi^*_{i,\lambda})_{i\in I}$ with $\pi^*_{i,\lambda} = \pi^*_i|_{T^*_{i,\lambda}}$.[33]

[33] Following Mertens and Zamir (1985) and Brandenburger and Dekel (1993), but replacing Kolmogorov's extension theorem by a version generalized to signed measures with uniformly bounded total variations, we can identify $T^*_{i,\lambda}$

B.8 Rationalizability and Strategic Distinguishability

We define the set of λ-rationalizable actions, denoted by $R_{i,\lambda}(t_i)$ or $R_{i,\lambda}(t_i;\mathcal{T},\mathcal{M})$, as follows:

$$
\begin{aligned}
R^0_{i,\lambda}(t_i) &= M_i, \\
R^{n+1}_{i,\lambda}(t_i) &= \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \succsim_i \in P_\lambda(M_{-i}\times T_{-i}) \text{ s.t.} \\ \text{(i) } \succsim_i \text{ is certain of the graph of } R^n_{-i,\lambda}, \\ \text{(ii) } \operatorname{mrg}_{T_{-i}} \succsim_i = \pi_i(t_i), \\ \text{(iii) } \succsim_i \text{ weakly prefers } O(\cdot \mid m_i,\cdot) \text{ to } O(\cdot \mid m_i',\cdot) \\ \qquad \text{for any } m_i' \in M_i \end{array} \right\}, \\
R_{i,\lambda}(t_i) &= \bigcap_{n=0}^{\infty} R^n_{i,\lambda}(t_i).
\end{aligned}
$$

Let $d^*_{i,\lambda}$ be a metric compatible with the product topology on $T^*_{i,\lambda}$.

Proposition 11 *Fix $\lambda > 0$. For every $\varepsilon > 0$, there exists a mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$ such that*

$$
d^*_{i,\lambda}(\hat{\pi}_i(t_i;\mathcal{T}),\hat{\pi}_i(t_i';\mathcal{T}')) > \varepsilon \Rightarrow R_{i,\lambda}(t_i;\mathcal{T},\mathcal{M}) \cap R_{i,\lambda}(t_i';\mathcal{T}',\mathcal{M}) = \emptyset
$$

for any two λ-continuous type spaces $\mathcal{T} = (T_i,\pi_i)_{i\in I}$ and $\mathcal{T}' = (T_i',\pi_i')_{i\in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t_i' \in T_i'$.

Sketch of the Proof. The proof is analogous to that of Proposition 2. By Proposition 9, $H^{|I|-1}_{\lambda,n-1}$ is compact and metrizable, and hence we can let $X = H^{|I|-1}_{\lambda,n-1}$ and apply Lemma 6 repeatedly. ■

Together with the fact that equilibrium is a refinement of λ-rationalizability, Proposition 11 implies Proposition 6. Indeed, Propositions 4 and 11 imply the following characterization of strategic distinguishability for λ-rationalizability.

with the set of all coherent λ-continuous preference hierarchies. Thus $T^*_{i,\lambda}$ is compact and metrizable, and $\pi^*_{i,\lambda}\colon T^*_{i,\lambda} \to P_\lambda(T^*_{-i,\lambda})$ is a homeomorphism. We do not need these facts, though.

Theorem 3 *Fix $\lambda > 0$. For any two λ-continuous countable type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T'_i, \pi'_i)_{i \in I}$, any agent $i \in I$, and any two types $t_i \in T_i$ and $t'_i \in T'_i$, the following three conditions are equivalent:*

1. *$\hat{\pi}_i(t_i; \mathcal{T}) = \hat{\pi}_i(t'_i; \mathcal{T}')$;*
2. *$E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t'_i; \mathcal{T}', \mathcal{M}) \neq \emptyset$ for any mechanism $\mathcal{M}$.*
3. *$R_{i,\lambda}(t_i; \mathcal{T}, \mathcal{M}) \cap R_{i,\lambda}(t'_i; \mathcal{T}', \mathcal{M}) \neq \emptyset$ for any mechanism $\mathcal{M}$.*

Proof. $1 \Rightarrow 2$ follows from Proposition 4. $2 \Rightarrow 3$ follows from the fact that equilibrium is a refinement of λ-rationalizability. $3 \Rightarrow 1$ follows from Proposition 11. ■

References

Abreu, D., and H. Matsushima (1992a): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

——— (1992b): "Virtual Implementation In Iteratively Undominated Strategies: Incomplete Information," Discussion paper, Princeton University and University of Tokyo.

Afriat, S. (1967): "The Construction of Utility Functions from Expenditure Data," *International Economic Review*, 8, 67–77.

Battigalli, P., and M. Dufwenberg (2009): "Dynamic Psychological Games," *Journal of Economic Theory*, 144, 1–35.

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

Battigalli, P., A. D. Tillio, E. Grillo, and A. Penta (2011): "Interactive Epistemology and Solution Concepts for Games with Asymmetric Information," *B E Press Journal of Theoretical Economics: Advances*, 1, Article 6.

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

——— (2012): *Robust Mechanism Design*. World Scientific Publishing, Singapore.

Bergemann, D., S. Morris, and S. Takahashi (2011): "Interdependent Preferences and Strategic Distinguishability," Discussion paper, Princeton University and Yale University.

Brandenburger, A., and E. Dekel (1993): "Hierarchies of Belief and Common Knowledge," *Journal of Economic Theory*, 59, 189–198.

Brooks, B. (2014): "Extracting Common Knowledge: Strengthening a Folk Argument," Discussion paper, Princeton University and University of Chicago.

Chambers, C. (2008): "Proper Scoring Rules for General Decision Models," *Games and Economic Behavior*, 63, 32–40.

Chambers, C., and N. Lambert (2014): "Dynamically Eliciting Unobservable Information," Discussion paper, University of California at San Diego and Stanford University.

Dasgupta, P., and E. Maskin (2000): "Efficient Auctions," *Quarterly Journal of Economics*, 115, 341–388.

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

——— (2007): "Interim Correlated Rationalizability," *Theoretical Economics*, 2, 15–40.

di Tillio, A. (2008): "Subjective Expected Utility in Games," *Theoretical Economics*, 3, 287–323.

Ely, J. C., and M. Peski (2006): "Hierarchies of Belief and Interim Rationalizability," *Theoretical Economics*, 1, 19–65.

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

Ganguli, J., and A. Heifetz (2012): "Universal Interactive Preferences," Discussion paper, University of Essex.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1, 60–79.

Gossner, O., and J. Mertens (2001): "The Value of Information in Zero-Sum Games," Universite Paris-Nanterre and CORE, University Catholique de Louvain.

Grant, S., I. Meneghel, and R. Tourky (2014): "Savage Games," Discussion paper, The University of Queensland and Australian National University.

Gul, F., and W. Pesendorfer (2010): "Interdependent Preference Models as a Theory of Intentions," Discussion paper, Princeton University.

Heifetz, A., and Z. Neeman (2006): "On the Generic (Im)Possibility of Full Surplus Extraction in Mechanism Design," *Econometrica*, 74, 213–233.

Heifetz, A., and D. Samet (1998): "Typology-Free Typology of Beliefs," *Journal of Economic Theory*, 82, 324–341.

Levine, D. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1, 593–622.

Mertens, J., and S. Zamir (1985): "Formalization of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 14, 1–29.

Morris, S., and S. Takahashi (2012): "Games in Preference Form and Preference Rationalizability," Discussion paper, Princeton University.

Nagata, J. (1985): *Modern General Topology*. North-Holland.

Perry, M., and P. Reny (2002): "An Ex Post Efficient Auction," *Econometrica*, 70, 1199–1212.

Sadzik, T. (2010): "Beliefs Revealed in Bayesian-Nash Equilibrium," Discussion paper, New York University.

Savage, L. (1954): *The Foundations of Statistics*. Wiley, New York, 1st edn.

Sprumont, Y. (2000): "On The Testable Implications of Collective Choice Theories," *Journal of Economic Theory*, 93, 205–232.