

# The Canonical Space for Behavioral Types<sup>†</sup>

Faruk Gul  
and  
Wolfgang Pesendorfer

Princeton University

July 2007

## Abstract

We develop a model of behavioral types to analyze situations in which an agent's preferences depend on others' characteristics and personalities. We define a canonical type space and provide conditions under which an abstract type space is a component of the canonical type space. As an application, we develop a model of reciprocity in which agents reward the kindness of others.

---

<sup>†</sup> This research was supported by grants SES9911177, SES0236882, SES9905178 and SES0214050 from the National Science Foundation. We thank Stephen Morris, Phil Reny, four anonymous referees and the editor for their comments.

## 1. Introduction

In many economic models, preference interdependencies arise from informational interdependencies. In a standard asymmetric information model, a type profile,  $t = (t_1, t_2)$ , yields a von Neumann-Morgenstern utility function,  $u_i(\cdot, t)$ , for each player. Thus, player 1's preference depends both on his type and the type of his opponent. In these models, types serve as (a part of) the description of each players' information (or knowledge) about some set of payoff relevant parameters.

In this paper, we develop a framework for modeling interdependent preferences to describe phenomena such as reciprocity, conformity, and spitefulness; that is, phenomena that arise not from interactive information but from behavioral concerns. A type in our model represents relevant personality attributes rather than information. These attributes determine both the person's and his opponents' behavior.

To distinguish between our types and the types in asymmetric information (or interactive knowledge) models we refer to the former as behavioral types and the latter as informational (or epistemic) types. Consider the following situation as an illustration of this distinction. Two agents must decide how to share a fixed quantity of goods. The possible social outcomes (denoted  $A$ ) allocate all goods to one person or share them evenly. Then,

$$A = \{(1, 0), (.5, .5), (0, 1)\}$$

The generous preference,  $G$ , prefers sharing while the selfish preference,  $S$ , prefers getting all the goods. Both preferences rank getting 0 as the worst outcome. Suppose there are three types: type 3 is the nicest type who exhibits the generous preference irrespective of the other person's type. Type 1 has the generous preference if his opponent is type 3 and the selfish preference otherwise. Finally, type 2 is an intermediate type who has the generous preference unless his opponent is type 1.

Table 1 below summarizes this description. Each row corresponds to a player 1 type, each column to a player 2 type and a type profile determines a preference profile. The first

row describes type 1 (of player 1) by stating how he responds to player 2’s type; that is, player 1’s types are depicted as functions from player 2’s types to preference profiles.

	1	2	3
1	$(S, S)$	$(S, S)$	$(G, G)$
2	$(S, S)$	$(G, G)$	$(G, G)$
3	$(G, G)$	$(G, G)$	$(G, G)$

Table 1

We refer to a description such as the one given in Table 1 as an *Interdependent Preference Model* (IPM). Like asymmetric information models, IPMs are reduced-form descriptions of preference interdependence. Unlike asymmetric information models, IPMs do not describe information (or knowledge). Table 1 is silent as to whether player 1 knows player 2’s type; it only describes how each type will respond to opponent types. Therefore, Mertens-Zamir and Brandenburger-Dekel belief hierarchies or similar epistemic constructions cannot provide a suitable interpretation of behavioral types.<sup>1</sup> Our two main theorems offer an interpretation of behavioral types as *hierarchies of preference responses*. These hierarchies describe how the agent reacts (or would react) to others’ preference responses.

More formally, an IPM is a triple  $M = (T, \Gamma, \omega)$  where  $T$  is a compact set of behavioral types,  $\Gamma$  is a continuous function that assigns the preference profile  $\Gamma(t, t')$  to each type profile  $(t, t')$ , and  $\omega(t)$  is type  $t$ ’s characteristic. The preference profile  $(R_1, R_2) = \Gamma(t_1, t_2)$  specifies agent  $i$ ’s ranking of a fixed set,  $A$ , of physical consequences. Hence, the same IPM  $M$  is applicable in any context that has the same two agents 1, 2 and the same set of physical consequences. For example, whenever the same two agents 1 and 2 are playing a two-person normal form game or a two-person extensive form game or are in a two person competitive economy with consequences in  $A$ , their preferences will be  $(R_1, R_2)$ . Once the behavioral types are specified all of the “psychology” is resolved and we are left with two standard preferences.

A type’s characteristic are those attributes that can be understood without reference to behavior (i.e., preferences). Examples are physical attributes, social affiliations, and

---

<sup>1</sup> Behavioral types are akin to what Battigalli and Siniscalchi (2003) and Bergemann and Morris (2007) call “payoff types,” and an informational type is to be construed as a belief over the parameter space of behavioral types.

the position within a well-defined hierarchy.<sup>2</sup> We refer to the residual – those preference-relevant attributes that cannot be associated to a characteristic – as *personality*. Hence, the personality of a player refers to those attributes that can be described only in terms of preferences.

To complete the description of the IPM in Table 1, we must specify a characteristic for each type. If each type has a distinct characteristic (type 1 is a child, type 2 is a teenager and type 3 is an adult), then the labels 1, 2 and 3 require no further explanation since each type can be identified with an age group. However, if the types share a common characteristic and differ only in their personality, then the IPM’s description of a type is circular: agent 1’s personality is interpreted as a response to agent 2’s personality and vice versa. The preference response hierarchies developed in this paper resolve this circularity and describe types solely through preference statements; that is, they make no reference to the labels 1, 2, and 3 and therefore provide an explanation or interpretation of the three personalities.

Theorem 1 shows that every consistent collection of preference response hierarchies constitutes an IPM. That is, it ensures that each preference response hierarchy corresponds to a behavioral type and constructs the canonical space of behavioral types. Theorem 2 provides the converse by identifying a simple condition on an IPM that is necessary and sufficient for each behavioral type to be equivalent to a preference response hierarchy. To understand this condition – *validity* – consider any partition of the type space  $T$  such that some partition elements are not a singleton set. Validity fails if all types in any particular partition element have the same characteristic and for any two partition elements  $D, D'$ , types  $t_1, t_2, \in D$  and  $t_3 \in D'$ , there exists  $t_4 \in D'$  such that  $(t_2, t_4)$  yield the same preference profile as  $(t_1, t_3)$ . If no such partition exists, then the model is valid. Hence, validity requires that there be no partition of the type space such that types in each partition element are indistinguishable.

When a model fails validity, types cannot be expressed as “preference statements.” To formalize this observation, we develop a notion of communicability *in a given language*. We show that players can communicate their type in the *language of preferences* if and only

---

<sup>2</sup> This corresponds to the payoff types in Battigalli and Siniscalchi (2003) and Bergemann and Morris (2007).

if the model is valid (Theorem 3). Hence, invalid models have types with the same characteristic that are indistinguishable through behavior. Since we associate differences that can be identified without reference to behavior with different characteristics, we conclude that an invalid model is ill-defined.

As an application of our model, we present a definition of reciprocity and identify a class of valid interdependent preference models in which types reciprocate. Let  $A = [0, 1] \times [0, 1]$  denote the possible outcomes. For  $(a_1, a_2) \in A$ , the quantity  $a_1$  is the individual's own consumption and  $a_2$  is the opponent's consumption. Assume that a single parameter,  $r \in [\underline{r}, \bar{r}]$ , describes a preferences where  $r$  is the weight the agent puts on the opponent's payoff.<sup>3</sup> The preference  $r$  maximizes

$$ru(a_2) + (1 - r)u(a_1)$$

for some fixed utility function  $u$  and depends on the player's and his opponent's type; in particular, let  $\gamma(t, t') \in [\underline{r}, \bar{r}]$  denote the preference of type  $t$  when the opponent is type  $t'$ . A higher  $r$  means a *kinder* preference. Type  $t$  is *nicer than* type  $t'$  if  $t$  is kinder than  $t'$  to every type of the opponent; that is,  $\gamma(t, \cdot) \geq \gamma(t', \cdot)$ . A type *reciprocates* if it is kinder to a nicer opponent, i.e.,  $\hat{t}$  reciprocates if  $\gamma(\hat{t}, t) \geq \gamma(\hat{t}, t')$  whenever  $t$  is nicer than  $t'$ . An example of reciprocating types are *symmetric* IPMs such that  $\gamma(t, t') = \gamma(t', t)$ . Symmetry is a strong form of reciprocity where “no type is kinder than his opponent.”

For the case of two possible preferences, we show that in all valid symmetric IPMs types can be ordered by how nice they are. This leads to a simple characterization of the resulting interdependent preference models, provided in Theorem 4. We also characterize all valid IPMs for the case in which the set of possible preferences is an interval and types can be ordered according to niceness (Theorem 5). Levine (1998) provides a reciprocity model that is a special case of the class of IPMs studied in Theorem 5.

Our model and, in particular, its application to reciprocity is related to work on psychological games (Geanakoplos, Pearce and Stacchetti (1989)) where players' preferences may depend directly on other players' beliefs. Several authors<sup>4</sup> have proposed models

---

<sup>3</sup> Note that  $r$  may be negative. A preference  $r < 0$  is spiteful, i.e., the individual is willing to trade off a reduction in his own reward for a reduction in the opponent's reward.

<sup>4</sup> See Sobel (2004) for a survey. See also Charness and Rabin (2002), Cox and Friedman (2002), Falk and Fischbacher (1998) for other models related to GPS.

of reciprocity based on GPS. Rabin (1993) develops a theory of fairness and reciprocity in normal form games. Dufwenberg and Kirchsteiger (2005) propose such a theory for extensive form games. Segal and Sobel (2003) assume that a player’s utility function is parameterized by the [player’s belief about the] strategy profile in the game. They provide axioms yielding utility functions that are separable in the player’s own and his opponents “selfish utility”.

GPS-based models and our IPMs both deal with preference interdependence. However, the two approaches have different emphasis and different advantages, and therefore are likely to be appropriate in different contexts. For example, psychological games can describe situations in which player 2 has no strategic choice yet player 1 derives utility from surprising him. This would be difficult to do in our framework since we designate personalities and characteristics, rather than beliefs, as the carriers of utility. The two approaches also have important similarities. With our approach, when there is uncertainty about the opponents behavioral type, beliefs regarding strategies will often be related to the underlying beliefs about personalities. This dependence ensures that beliefs over strategies have payoff consequences beyond what they imply about the physical outcome of the game. Hence, when payoffs depend on physical consequences and personalities and, in addition, there is asymmetric information about personalities, the model replicates many (but not all) of the effects of a model that postulates a “direct” dependence of utilities on beliefs (about the opponent’s action).

An advantage of our approach is that, like standard game theory, it distinguishes between context-specific incentives (games, budget-sets, contracts) and context-independent payoffs (preferences, personalities) and therefore can be used to analyze both strategic and competitive situations.<sup>5</sup> Indeed, with our framework, it is easy to model a situation in which the same group of agents (i.e., behavioral types) play different games and/or appear in different competitive economies. Incorporating behavioral concerns into normal form or extensive form games or competitive economies requires no new ideas or equilibrium concepts. With GPS-based models, additional hypotheses and novel constructions are needed to identify players in different games as being the same person. Developing extensive form

---

<sup>5</sup> This separation is central to implementation theory.

psychological games requires new constructions.<sup>6</sup> Similarly, new constructions are needed to analyze “psychological economies” or “psychological implementation theory.”

Our canonical type space provides a foundation for valid IPMs that is analogous to the Mertens and Zamir (1985) and Brandenburger and Dekel (1993) foundations for informational (Harsanyi) types. In the concluding section of the paper we provide a detailed discussion of the relation to this literature and, in particular, to Mariotti, Meier and Piccione (2004), Bergemann and Morris (2007) and the literature on communication and consensus (Geanakoplos and Polemarchakis (1982), Cave (1983), Bacharach (1985), Parikh and Krasucki (1990)). All proofs are in the appendix.

## 2. Behavioral Types

### 2.1 Preliminaries

Let  $A$  denote the compact metric space of alternatives. A binary relation  $R$  on  $A$  is *transitive* if  $xRy, yRz$  implies  $xRz$  for all  $x, y, z \in A$ . The binary relation  $R$  is *complete* if either  $xRy$  or  $yRx$  holds for all  $x, y \in A$ . If  $R$  is both transitive and complete, we say that  $R$  is a preference relation. The binary relation  $R$  is *continuous* if for all  $x \in A$ , the sets  $\{y \in A \mid yRx\}, \{y \in A \mid xRy\}$  are closed subsets of  $A$ .

When  $X_j$  is a metric space for all  $j$  in some countable or finite index set  $J$ , we endow  $\times_{j \in J} X_j$  with the sup metric. For any compact metric space  $X$ , let  $\mathcal{H}_X$  be the set of all nonempty, closed subsets of  $X$  and endow  $\mathcal{H}_X$  with the Hausdorff topology. For the compact metric spaces  $X, Z$  let  $\mathcal{C}(X, \mathcal{H}_Z)$  denote the set of all functions  $f : X \rightarrow \mathcal{H}_Z$  such that their graph  $G(f) = \{(x, z) \in X \times Z \mid z \in f(x)\}$  is closed in  $X \times Z$ .<sup>7</sup> We endow  $\mathcal{C}(X, \mathcal{H}_Z)$  with the following metric:  $d(f, g) = d_H(G(f), G(g))$ , where  $d_H$  is the Hausdorff metric on the set of all nonempty closed subsets of  $X \times Z$ . We identify the function  $f : X \rightarrow Z$  with the function  $\bar{f} : X \rightarrow \mathcal{H}_Z$  such that  $\bar{f}(x) = \{f(x)\}$  for all  $x \in X$ . It is easy to verify that such a function  $f$  is an element of  $\mathcal{C}(X, \mathcal{H}_Z)$  if and only if  $f$  is continuous. We use  $\mathcal{C}(X, Z) \subset \mathcal{C}(X, \mathcal{H}_Z)$  to denote the set of continuous functions from  $X$  to  $Z$ .

---

<sup>6</sup> See Battigalli and Dufwenberg (2005)’s definition of sequential equilibrium for extensive form psychological games.

<sup>7</sup> Hence,  $\mathcal{C}(X, \mathcal{H}_Z)$  is the set of upper hemi-continuous correspondences from  $X$  to  $Z$ .

## 2.2 Interdependent Preference Models

We model the social environment by assuming that there is one individual other than the decision-maker whose type affects the behavior of the decision-maker. The two-person setting simplifies the notation and the extension to more than two agents is straightforward. An interdependent preference model (IPM) associates to each pair of types a preference profile and, in addition, specifies the characteristic of every type. A characteristic refers to a physical attribute of a type such as the type's ethnicity or gender.

Recall that  $A$  as is a compact metric space of alternatives. Let  $\mathcal{R} \subset \mathcal{H}_{A \times A}$  be a nonempty and compact set of continuous preference relations on  $A$ . The compact metric spaces  $T$  and  $\Omega$  are the type space and the set of possible characteristics, respectively.

**Definition:** Let  $\gamma : T \times T \rightarrow \mathcal{R}$  and  $\omega : T \rightarrow \Omega$  be a continuous functions. Then,  $M = (T, \gamma, \omega)$  is an interdependent preference model (IPM).

For the IPM  $(T, \gamma, \omega)$  the type profile  $(t, t')$  implies the preference profile

$$\Gamma(t, t') := (\gamma(t, t'), \gamma(t', t)) \tag{1}$$

Below, we sometimes refer to an IPM  $M = (T, \Gamma, \omega)$ . In that case, it is understood that  $\Gamma$  satisfies (1) for some  $\gamma : T \times T \rightarrow \mathcal{R}$ .

We use the concept of a characteristic to identify those aspects of a type that correspond to readily identifiable differences between agents. The inclusion of a characteristic renders the assumed symmetry of the function  $\Gamma$  without loss of generality since the characteristic can be used to identify a player's role. When two types share the same characteristic ( $\omega(t) = \omega(t')$ ) but  $\gamma(t, \cdot) \neq \gamma(t', \cdot)$  then we say that types  $t$  and  $t'$  differ in their *personality*.

**Example 1:** A fixed sum of money can either be shared equally between the players or given to one of the players. There are two possible preferences for each player; either the player may have a selfish preference (denoted  $S$ ) that ranks getting all of the money above sharing it equally or the player may have a generous preference (denoted  $G$ ) that ranks sharing the surplus above getting all of the surplus. Giving all the money to the opponent

is always the least preferred option. There are  $k$  types who share the same characteristic and

$$\gamma(i, j) = \begin{cases} G & \text{if } i + j > k \\ S & \text{if } i + j \leq k \end{cases}$$

For  $k = 3$ , type 3 is the most generous type who prefers the generous distribution irrespective of the opponent type. Type 1 is the least generous type who prefers the generous distribution only if the opponent is the most generous type. Type 2 is of intermediate generosity and prefers the generous distribution unless the opponent is the least generous type.

In Example 1, the agent's personality identifies his type; in particular, his level of generosity. However, as is clear from the example, the "type" of a player is an arbitrary label and not a plausible primitive of the model. To interpret the types in Example 1, we associate with each type preference statements of the form: "*type 3 always prefers the generous distribution*" or "*type 1 always prefers the selfish distribution unless the opponent always prefers the generous distribution*". In the following section, we will develop a canonical type space based on such preference statements.

### 2.3 The Canonical Type Space

The canonical type space depends on the set of alternatives,  $A$ , the set of preferences  $\mathcal{R}$  and the space of characteristics  $\Omega$ . For notational simplicity we suppress the dependence of the canonical type space on the compact metric spaces  $A, \mathcal{R}, \Omega$ . Let  $\mathcal{H} = \mathcal{H}_{\mathcal{R} \times \mathcal{R}}$  denote the collection of non-empty, closed subsets of the set of preference profiles.

We define a sequence of sets that represent a *system of preference response hierarchies*:

**Definition:** *A collection of nonempty compact sets  $(\Theta_0, \Theta_1, \dots)$  is a system of preference response hierarchies if  $\Theta_0 = \Omega$  and*

$$\Theta_n \subset \Theta_{n-1} \times \mathcal{C}(\Theta_{n-1}, \mathcal{H})$$

for all  $n \geq 1$ .

The entry  $\theta_0 \in \Theta_0$  specifies a characteristic. The entry  $\theta_1$  specifies a characteristic and a map that associates each opponent characteristic with a set of preference profiles.

More generally, the entry  $\theta_k$  consists of the previous entry ( $\theta_{k-1}$ ) and the function  $f_k$  that specifies for each  $\theta_{k-1}$  (of the opponent) a set of possible preference profiles.

**Example 1:** (Representation of types). We can interpret preference response hierarchies as a sequence of reports on the set of possible preference profiles for each type. In response to the opponent's reports, each type gradually eliminates preferences from the set of possible profiles until a unique preference profile emerges for each opponent type. In Example 1,  $\gamma(i, j) = \gamma(j, i)$  and therefore we can identify each preference profile with player 1's preference. All types have the same characteristic and hence there is nothing to report in round 0. The set of possible preferences for types  $1, \dots, k-1$  are  $\{S, G\}$  and therefore the round 1 statement is  $\{S, G\}$  for those types. Type  $k$  always prefers the generous distribution and therefore the round 1 report of  $k$  is  $\{G\}$ . We conclude that

$$\Theta_1 = \{\{S, G\}, \{G\}\}$$

and note that round 1 identifies type  $k$ , the most generous type. In round 2, each type reports the set of possible preference profiles for every possible round 1 statement of the opponent. For type 1,

$$\theta_2(\{S, G\}) = \{S\}, \theta_2(\{G\}) = \{G\}$$

whereas for types  $2, \dots, k-1$ ,

$$\theta_2(\{S, G\}) = \{S, G\}, \theta_2(\{G\}) = \{G\}$$

Notice that after round 2, type 1 has a unique preference for every possible report of the opponent. Therefore, round 2 identifies type 1, the least generous type. Continuing in this fashion, round 3 specifies the set of possible preferences for each possible statement of the opponent in the previous two rounds, i.e., for each pair  $(\theta_1, \theta_2)$ . It is easy to see that round 3 identifies type  $k-1$ , the second most generous type, and continuing in this fashion all types are identified in  $k-1$  rounds.

The entries  $\theta_n \in \Theta_n$  must satisfy two consistency requirements: first, if  $\theta_{n+1} = (f_0, \dots, f_{n+1}) \in \Theta_{n+1}$ , then  $f_{i+1}$  and  $f_i$  must be consistent for all  $i \leq n$ , that is,  $f_n(\theta'_{n-1})$

must be the union of the sets  $f_{n+1}(\theta'_n)$  taken over all the possible continuations  $\theta'_n$  of  $\theta'_{n-1}$ . Second, the round  $n$  statements of  $\theta_n = (\theta'_{n-1}, f_n)$  and  $\theta'_n = (\theta'_{n-1}, f'_n)$  must be compatible, that is,  $f_n(\theta'_{n-1})$  must contain a preference profile  $(R, R')$  such that  $(R', R)$  is contained in  $f'_n(\theta_{n-1})$ .

**Definition:** *The system of preference response hierarchies  $(\Theta_0, \Theta_1, \dots)$  is consistent if for all  $n \geq 1$  and  $(\theta_{n-1}, f_n, f_{n+1}) \in \Theta_{n+1}$ ,*

$$(i) f_n(\theta'_{n-1}) = \bigcup_{\{f'_n \mid (\theta'_{n-1}, f'_n) \in \Theta_n\}} f_{n+1}(\theta'_{n-1}, f'_n) \text{ for all } \theta'_{n-1} \in \Theta_{n-1}.$$

(ii) For all  $(\theta'_{n-1}, f'_n) \in \Theta_n$  there is  $(R, R') \in \mathcal{R} \times \mathcal{R}$  such that  $(R, R') \in f_n(\theta'_{n-1})$  and  $(R', R) \in f'_n(\theta_{n-1})$ .

For a given consistent system of preference response hierarchies  $(\Theta_0, \Theta_1, \dots)$ , we define a type to be a sequence  $(f_0, f_1, \dots)$  with the property that  $(f_0, \dots, f_n) \in \Theta_n$ . To qualify as a component of the canonical type space,  $\Theta$  must satisfy an additional property. Every type must generate a unique preference when confronted with any other type in the component  $\Theta$ . This means that for every pair of types  $(f_0, f_1, \dots), (f'_0, f'_1, \dots)$  it must be the case that  $f_n(f'_0, \dots, f'_{n-1})$  converges to a singleton as  $n \rightarrow \infty$ . Let  $\theta(n) = (f_0, f_1, \dots, f_n)$  denote the  $n$ -truncation of the sequence  $\theta = (f_0, f_1, \dots)$ .

**Definition:** *Let  $(\Theta_0, \Theta_1, \dots)$  be a consistent sequence of preference response hierarchies. Let  $\Theta := \{\theta \in \Theta_0 \times \prod_{n=1}^{\infty} \mathcal{C}(\Theta_{n-1}, \mathcal{H}) \mid \theta(n) \in \Theta_n\}$ . Then  $\Theta$  is a component of behavioral types if  $\Theta$  is compact and for all  $\theta = (f_0, f_1, \dots) \in \Theta$*

$$\bigcap_{n \geq 0} f_{n+1}(\theta'(n)) \text{ is a singleton}$$

for all  $\theta' \in \Theta$ .

The canonical type space is the union of all the components of behavioral types. Let  $\mathcal{I}$  denote the set of all components of interdependent types. The set

$$\mathcal{F} = \bigcup_{\Theta \in \mathcal{I}} \Theta$$

is the *canonical behavioral type space* or simply the canonical type space. Note that each element  $\theta \in \mathcal{F}$  belongs to a unique component  $\Theta \in \mathcal{I}$ . Hence,  $\mathcal{I}$  is a decomposition (or partition) of  $\mathcal{F}$ .

For any  $\Theta \in \mathcal{I}$ , let  $\Psi : \Theta \times \Theta \rightarrow \mathcal{S}$  denote the function that specifies a preference profile when the player is type  $\theta$  and the opponent is type  $\theta'$ . Hence,

$$\Psi(\theta, \theta') := \bigcap_{n \geq 0} f_{n+1}(\theta'(n)) \text{ for } (f_0, f_1, \dots) = \theta$$

Requirement (ii) in the definition of consistency ensures that the function  $\psi$  satisfies the following symmetry condition.

$$\Psi(\theta, \theta') = (R, R') \text{ implies } \Psi(\theta', \theta) = (R', R) \quad (S)$$

If  $\Psi$  satisfies (S), we say that  $\Psi$  is *symmetric*. We define  $\phi : \Theta \rightarrow \Omega \times \mathcal{C}(\Theta, \mathcal{S})$  to be the function that specifies for every type  $\theta \in \Theta$  the characteristic of the type  $\theta$  and the mapping  $\theta$  uses to assign preferences profile to opponent types. Hence,

$$\phi(\theta) := (f_0, \Psi(\theta, \cdot))$$

**Theorem 1:** *The function  $\Psi$  is continuous and symmetric and  $\phi$  is a homeomorphism from  $\Theta$  to  $\phi(\Theta)$ .*

An immediate consequence of Theorem 1 is that any component  $\Theta \in \mathcal{I}$  is an interdependent preference model as defined above. For a symmetric  $\Psi$ , there is a  $\psi : \Theta \times \Theta \rightarrow \mathcal{R}$  such that

$$\Psi(\theta, \theta') = (\psi(\theta, \theta'), \psi(\theta', \theta))$$

**Corollary:** *Let  $\Theta$  be a component of the canonical type space. Then  $M^\Theta = (\Theta, \psi, \omega)$  is an IPM.*

Note that  $\Theta$  is compact (by definition) and  $\psi$  is continuous by Theorem 1. Therefore, it follows that  $M^\Theta$  is an IPM. Theorem 2 in section 4 provides a converse to the Corollary above. It characterizes all those IPM's that correspond to some component of the canonical type space.

## 2.4 Valid Models

Not every IPM represents a component of the canonical type space. To qualify as a component of the canonical type space, types must be uniquely identified by the hierarchy of preference statements described in the previous section.

A decomposition (partition)  $\mathcal{D}$  of  $T$  is a pairwise disjoint collection of non-empty subsets with  $\bigcup_{D \in \mathcal{D}} D = T$ . Let  $D^t$  denote the unique element of  $\mathcal{D}$  that contains  $t$ . The decomposition  $\mathcal{D} = \{\{t\} \mid t \in T\}$  is called the *finest* decomposition. Let  $M = (T, \gamma, \omega)$  be an IPM. Let

$$\Gamma(t, t') := (\gamma(t, t'), \gamma(t', t))$$

and  $\Gamma(t, D) := \{\Gamma(t, t') \mid t' \in D\}$ .

**Definition:** *The decomposition  $\mathcal{D}$  challenges the IPM  $M = (T, \gamma, \omega)$  if it is not the finest decomposition and*

$$(i) \ t, t' \in D \in \mathcal{D} \text{ implies } \omega(t) = \omega(t')$$

$$(ii) \ t' \in D^t \in \mathcal{D} \text{ implies } \Gamma(t, D) = \Gamma(t', D) \text{ for all } D \in \mathcal{D}.$$

*If no decomposition challenges  $M$  then  $M$  is valid.*

Validity requires that we cannot find a decomposition of the type space such that in some set there are multiple indistinguishable types. Theorem 2 shows that any valid IPM corresponds to a component  $\Theta \in \mathcal{I}$ . Two IPM's  $M = (T, \gamma, \omega)$ ,  $M' = (T', \gamma', \omega')$  are isomorphic if there exists a homeomorphism  $\iota : T \rightarrow T'$  such that  $\omega(t) = \omega'(\iota(t))$  and  $\gamma(s, t) = \gamma'(\iota(s), \iota(t))$  for all  $s, t \in T$ .

**Theorem 2:** *An interdependent preference model  $M = (T, \gamma, \omega)$  is valid if and only if it is isomorphic to a component of the canonical type space.*

**Example 1:** (Valid IPM) The IPM in Example 1 above is valid. To see this, consider any decomposition  $\mathcal{D}$  of  $\{1, \dots, k\}$  that challenges the IPM. For  $D \in \mathcal{D}$ , let  $s(D) = \max\{t - t' \mid t, t' \in D\}$  and choose  $D^* \in \mathcal{D}$  such that  $s(D^*) \geq s(D)$  for all  $D \in \mathcal{D}$ . Let  $\underline{t} = \min D^*$  and  $\bar{t} = \max D^*$ . Since  $\mathcal{D}$  is a challenging decomposition,  $s(D^*) = \bar{t} - \underline{t} > 0$ . Let  $t' = k - \underline{t}$  and note that  $t' > 0$ . Let  $D^{t'}$  be the element of  $\mathcal{D}$  that contains  $t'$ . Since  $\gamma(\underline{t}, t') = S$  and

$\gamma(\bar{t}, t') = G$ , we conclude that  $\gamma(\underline{t}, D^{t'}) = \gamma(\bar{t}, D^{t'}) = \{G, S\}$ . Hence,  $\min D^{t'} \leq k - \bar{t}$  and  $\max D^{t'} \geq k + 1 - \underline{t}$  and therefore,  $s(D^{t'}) = \max D^{t'} - \min D^{t'} \geq \bar{t} - \underline{t} + 1 = s(D^*) + 1$ , a contradiction.

**Example 2:** (Invalid IPM). Consider the following modified version of example 1. Let  $r \in \{1, \dots, k\}$  with  $r > (k + 1)/2$  and define

$$\gamma(i, j) = \begin{cases} G & \text{if } i + j > k, i \neq r \text{ or } j \neq r \\ S & \text{if } i = j = r \\ S & \text{if } i + j \leq k \end{cases}$$

Example 2 is identical to Example 1 except that  $\gamma(r, r)$  is changed from  $G$  to  $S$ . To verify that Example 2 is not valid, let  $\mathcal{D} = \{D^t \mid t \in T\}$ , where  $D^t = \{t\}$  for  $t \notin \{k - r + 1, \dots, r\}$  and  $D^r = \{k - r + 1, \dots, r\}$ . It is straightforward to verify that part (ii) of the above definition is satisfied for this decomposition and therefore the model is not valid.

Validity captures the idea that the process of refining the set of possible types through preference statements does not terminate until the finest decomposition is reached. In Example 2, this process stops at a decomposition in which the types in  $D^r$  remain indistinguishable. In the next section, we demonstrate that in non-valid IPM's players cannot communicate their type with a language that is confined to preference statements.

### 3. Validity and Communicability

In the previous section, we informally interpreted the hierarchies as descriptions of a communication between players to determine the appropriate preference profile. In this section, we provide a definition of communicability and show that an IPM is communicable if and only if it is valid.

To analyze communication, we consider an epistemic model where each player knows his own behavioral type but does not know the opponent's behavioral type. The epistemic model consists of a finite set of states  $S$ , a map  $\nu : S \rightarrow \mathcal{R} \times \mathcal{R}$  that associates a preference profile to each state and a pair of decompositions of  $S$ , denoted  $\mathcal{T}_i$ , that represent the knowledge of players  $i = 1, 2$ . A set  $A \in \mathcal{T}_i$  represents player  $i$ 's knowledge at state  $s \in A$  and hence the elements of the decomposition  $\mathcal{T}_i$  are the epistemic types for player  $i$ .

Let  $M = (T, \gamma)$  be an IPM with a single characteristic<sup>8</sup> and a finite set of types. The epistemic model  $E = \{S, \mathcal{T}_1, \mathcal{T}_2, \nu\}$  is *equivalent* to the IPM  $M = (T, \gamma)$  if there is a bijection  $\zeta_i : T \rightarrow \mathcal{T}_i$  such that  $\{\Gamma(t, t')\} = \nu(\zeta_1(t) \cap \zeta_2(t'))$  for all  $t, t' \in T$ . The bijection  $\zeta_i$  maps behavioral types (of  $M$ ) into epistemic types (of  $E$ ) while preserving the resulting preference profile. We refer to  $E$  as *an IPM in epistemic form (IPM-EF)*.

A collection of subsets  $\mathcal{K}$  of a set  $X$  is an algebra if it contains  $X$  and is closed under unions and complements. For any two algebras  $\mathcal{K}, \mathcal{L}$ , let  $\mathcal{K} \vee \mathcal{L}$  denote the smallest (in terms of set inclusion) algebra that contains both and let  $\mathcal{K} \wedge \mathcal{L}$  denote the largest algebra contained in both  $\mathcal{K}, \mathcal{L}$ .

Communicability of a model is defined with respect to a language that forms the basis of all communication. A *language*  $\mathcal{L}$  is an algebra on  $S$  and  $A \in \mathcal{L}$  is a *word*. For example,  $A \subset S$  is a word in the language of preferences if  $A = \{s \in S \mid \nu(s) \in V\}$  for some set of preferences  $V \in \mathcal{H}$ .

**Definition:** *The language of preferences is the algebra  $\{A \subset S \mid A = \nu^{-1}(V), V \in \mathcal{H}\}$ . The language of types is the smallest algebra containing the sets  $\mathcal{T}_1$  and  $\mathcal{T}_2$ .*

We illustrate the definitions by applying them to Example 1 from the previous section.

**Example 1:** (epistemic form) Let  $M = (T, \gamma)$  with  $T = \{1, \dots, k\}$  be the IPM described in Example 1 above. The following IPM-EF is equivalent to  $M$ . The state space  $S$  is given by  $S = \{(i, j) \mid i \in T, j \in T\}$ , the function  $\nu$  is given by  $\nu(i, j) := (\gamma(i, j), \gamma(j, i))$ , where

$$\gamma(i, j) = \begin{cases} G & \text{if } i + j > k \\ S & \text{if } i + j \leq k \end{cases}$$

(as in Example 1 above). The information partitions are  $\mathcal{T}_1 = \{\{(i, 1), \dots, (i, k)\} \mid i \in T\}$  and  $\mathcal{T}_2 = \{\{(1, i), \dots, (k, i)\} \mid i \in T\}$ . There are two possible preference profiles in this example,  $\{G, G\}$  and  $\{S, S\}$ . Therefore, the language of preferences has three non-empty words. These are  $A = \{(i, j) \mid i + j > k\}$ ,  $B = \{(i, j) \mid i + j \leq k\}$  and  $S = A \cup B$ . The word  $A$  corresponds to the set of preference profiles  $\{(G, G)\}$ , the word  $B$  corresponds to the set of preference profiles  $\{(S, S)\}$  and the word  $S = A \cup B$  corresponds to the set of preference

---

<sup>8</sup> Extending the analysis below to IPMs with multiple characteristics is straightforward. We assume a single characteristic to keep the notation simple.

profiles  $\{(S, S), (G, G)\}$ . By contrast, the language of types is the finest possible language. It contains every state  $\{(i, j)\}$  as a possible word.

A person with knowledge  $\mathcal{T}_i$  can utilize the word  $A \in \mathcal{L}$  to make a statement about whether or not he knows  $A$ . Let

$$\mathcal{T}_i * A := \bigcup_{B \in \mathcal{T}_i, B \subset A} B$$

Then,  $i$  can use the word  $A$  to communicate the words  $\{\mathcal{T}_i * A, \mathcal{T}_i * (S \setminus A)\}$  to  $j$ . These are words derived from  $A$  and  $S \setminus A$  that  $i$  *understands*; that is, at every  $s \in S$ ,  $i$  knows whether or not  $\mathcal{T}_i * A$  and  $\mathcal{T}_i * (S \setminus A)$  applies (i.e., is true); he knows whether or not he knows  $A$  and he knows whether or not he knows  $S \setminus A$ . Then, using standard logical operations he can also communicate other words such as  $[S \setminus (\mathcal{T}_i * A)] \cap [S \setminus (\mathcal{T}_i * (S \setminus A))]$ ; i.e., that he knows neither  $A$  nor  $S \setminus A$ .

The collection  $\mathcal{T}_i * \mathcal{L}$  is the smallest algebra that contains  $\mathcal{T}_i * A$  for all  $A \in \mathcal{L}$ . It is easy to verify that  $\mathcal{T}_i * \mathcal{L}$  is contained in any algebra that contains  $\mathcal{T}_i$ . That is, in any language,  $i$  can only communicate a coarsening of his knowledge.

After the initial round of communication, agents have access to the richer language  $\hat{\mathcal{L}} = \mathcal{L} \vee [\mathcal{T}_1 * \mathcal{L}] \vee [\mathcal{T}_2 * \mathcal{L}]$  which in turn can be refined through communication. Let,  $\mathcal{L}^1 = \mathcal{L}$  and define inductively

$$\mathcal{L}^{n+1} = F(\mathcal{L}^n) := \mathcal{L}^n \vee [\mathcal{T}_1 * \mathcal{L}^n] \vee [\mathcal{T}_2 * \mathcal{L}^n]$$

and note that the language gets refined until  $\mathcal{L}^n$  is a fixed-point of  $F$ . Let  $G(\mathcal{L})$  denote this fixed point, i.e.,  $G(\mathcal{L}) = \mathcal{L}^n$  such that  $F(\mathcal{L}^n) = \mathcal{L}^n$ . It is straightforward to show that  $G(\mathcal{L})$  is well-defined.<sup>9</sup>

---

<sup>9</sup> Let  $\Lambda$  be the set of all algebras on  $S$ . The set  $\Lambda$  is a lattice under the binary relation  $\subset$ . The set  $\Lambda_{\mathcal{L}} := \{\hat{\mathcal{L}} \in \Lambda \mid \mathcal{L} \subset \hat{\mathcal{L}}\}$  is a sublattice of  $\Lambda$  and  $F$  is an increasing function on  $\Lambda_{\mathcal{L}}$ ; that is,

$$\mathcal{L}' \subset \mathcal{L}'' \text{ implies } F(\mathcal{L}') \subset F(\mathcal{L}'')$$

Let  $\hat{\mathcal{L}}$  be any fixed-point of  $F$  on  $\Lambda_{\mathcal{L}}$ . Then,  $\mathcal{L} = \mathcal{L}^1 \subset \hat{\mathcal{L}}$  and by induction  $G(\mathcal{L}) \subset F(\hat{\mathcal{L}}) = \hat{\mathcal{L}}$ . It follows that  $G(\mathcal{L})$  is the smallest fixed-point of  $F$  in  $\Lambda_{\mathcal{L}}$ . This observation permits the following protocol invariance result: suppose that each round agents exchange some (but not necessarily all) of their information until they reach a situation in which they have nothing new to convey. Then, they will have communicated  $\mathcal{C}_{\mathcal{L}}$ .

Starting with the language  $\mathcal{L}$ , the algebra

$$\mathcal{C}_{\mathcal{L}} := [\mathcal{T}_1 * G(\mathcal{L})] \vee [\mathcal{T}_2 * G(\mathcal{L})]$$

is the most that players can communicate to each other. Hence, any collection of subsets,  $\mathcal{M}$ , of  $S$  can be communicated in language  $\mathcal{L}$  if and only if  $\mathcal{M} \subset \mathcal{C}_{\mathcal{L}}$ .

**Definition:** An IPM  $E = (S, \mathcal{T}_1, \mathcal{T}_2, \nu)$  in epistemic form is communicable in language  $\mathcal{L}$  if  $\mathcal{T}_i \subset \mathcal{C}_{\mathcal{L}}$  for  $i = 1, 2$ .

An IPM is communicable in a particular language if players can communicate their types in that language. Obviously, every IPM in epistemic form is communicable in the language of types. In other words, if players have an agreed upon label for each type, then every IPM in epistemic form is communicable. Theorem 3 below shows that when players communicate in the language of preferences, they can communicate their types if and only if the IPM is valid.

**Theorem 3:** A finite IPM in epistemic form is communicable in the language of preferences if and only if it is equivalent to a valid IPM.

Note that communicability of an IPM implies that the resulting preference profile is communicable as well. This follows since every pair of types generates a unique preference profile and hence communicability (of types) implies communicability of preference profiles.

Any IPM (valid or not) is communicable in the language of types. However, players have access to the language of types only if these types can be referred to directly. Our model distinguishes between *types* and *characteristics* to separate those aspects of a player that can be referred to without explanation (*characteristics*) from those aspects that require an explanation as preference statements (*personality*). Hence, a situation in which players have access to the language of types is more appropriately modeled as a situation in which *all types have distinct characteristics*. Generalizing Theorem 3 to allow for multiple characteristics is straightforward. That generalization would establish the equivalence of validity and communicability in the language of preferences and characteristics.

## 4. Reciprocity

Reciprocity describes a personality that is kinder to nicer opponents. We consider a setting in which preferences can be ordered by their “kindness,” i.e., preferences can be described by a single real number  $r \in \mathcal{R}_0$ . We assume that  $\mathcal{R}_0$  is compact and that higher values of  $r$  are *kinder* preferences. For example, let  $(x_1, x_2)$  denote the monetary payoff of players and let

$$U_r(x_1, x_2) = (1 - r)x_1 + rx_2$$

denote the utility function representing preference  $r \in \mathcal{R}_0 = [-1/4, 1/2]$ . For simplicity, we assume that all types have the same characteristic.

**Definition:** Let  $M = (T, \gamma)$  in which  $\gamma : T \times T \rightarrow \mathcal{R}_0$ .

(i) Type  $t$  is nicer than type  $t'$  if  $\gamma(t, t'') \geq \gamma(t', t'')$  for all  $t''$ . The IPM  $M$  is ordered if for all  $t, t' \in T$ ,  $t$  is nicer than  $t'$  or  $t'$  is nicer than  $t$ .

(ii) Type  $t$  reciprocates if  $\gamma(t, t') \geq \gamma(t, t'')$  when  $t'$  is nicer than  $t''$ .

(iii) The IPM  $M = \{T, \gamma\}$  is a reciprocity model if it is valid, ordered, and every type reciprocates.

Consider the following simple class of interdependent preference models: A *binary-symmetric* IPM has two possible preferences ( $\mathcal{R}_0 = \{0, 1\}$ ) and satisfies  $\gamma(i, j) = \gamma(j, i)$ . For any  $k, m = 1, 2, \dots$  and  $K = \{1, \dots, k\}$ , define  $\gamma_m : K \times K \rightarrow \{G, S\}$  as follows

$$\gamma_m(i, j) = \begin{cases} G & \text{if } i + j > m \\ S & \text{otherwise} \end{cases}$$

Let  $M_k^0 = (K, \gamma_k)$  and let  $M_k^1 = (K, \gamma_{k+1})$ . The IPM  $M_k^0$  corresponds to Example 1 used throughout the previous sections. In it, the nicest type (type  $k$ ) brings about the profile  $(G, G)$  when matched with any opponent type. In the IPM  $M_k^1$ , the meanest (least nice) type (type 1) brings about the profile  $(S, S)$  when matched with any opponent types. In both models types are ordered.

Theorem 4 below shows that  $M_k^0$  and  $M_k^1$  are the only valid binary-symmetric models.

**Theorem 4:** A binary symmetric model  $M$  is valid if and only if it is isomorphic to  $M_k^0$  or  $M_k^1$  for some  $k$ .

Since  $M_k^0, M_k^1$  are reciprocity models, Theorem 4 implies that any binary symmetric model is a reciprocity model. To gain intuition for Theorem 4, first note that, by compactness, a binary reciprocity model  $M$  must have a finite number of types  $\{1, \dots, m\}$ . If  $m = 1$ , there is nothing to prove. Suppose the result is true whenever  $m = k$  and let  $m = k + 1$ . Since there are only two preferences, validity ensures that  $\gamma(t, \cdot)$  is constant for some type  $t$ . Suppose this constant is 1 and without loss of generality let  $t = m + 1$ .

It is easy to check that the validity and symmetry of  $M$  imply that  $(\{1, \dots, k\}, \gamma)$  is valid and that there exists no  $t \leq k$  such that  $\gamma(t, t') = 1$  for all  $t' \leq k$ . Then, by the inductive hypothesis, the restriction of  $\gamma$  to  $\{1, \dots, k\}$  must be  $\gamma_k$ , implying that  $\gamma = \gamma_k$ .

Next, we consider general (non-binary, non-symmetric) reciprocity models. Theorem 5 establishes that when studying reciprocity, we can assume that  $T \subset \mathbb{R}$  and  $\gamma$  is weakly increasing in both arguments. A function  $\zeta : K \rightarrow \mathcal{C}(K, \mathbb{R})$  is strictly increasing if it is one-to-one and  $x \geq y$  implies  $\zeta(x)(w) \geq \zeta(y)(w)$  for all  $w \in T$ . For any function  $\gamma : K \times K \rightarrow \mathbb{R}$ , define  $\zeta_\gamma : K \rightarrow \mathcal{C}(K, \mathbb{R})$  as follows:  $\zeta_\gamma(x)(w) = \gamma(x, w)$ . Hence,  $\zeta_\gamma$  is strictly increasing if it is one-to-one and  $\gamma$  is weakly increasing in both arguments. Theorem 5 also shows that any IPM  $(T, \gamma)$  such that  $T \subset \mathbb{R}$  is a reciprocity model if  $\zeta_\gamma$  is strictly increasing.

**Theorem 5:** *The IPM  $(T, \gamma')$  is a reciprocity model if and only if it is isomorphic to some  $(K, \gamma)$ , where  $K \subset \mathbb{R}$  and  $\zeta_\gamma$  is strictly increasing.*

Let  $(K, \gamma)$  be an IPM such that  $K$  is a compact set of reals and  $\zeta_\gamma$  is strictly increasing. Hence, every type reciprocates. To see that  $(K, \gamma)$  is valid, consider a decomposition  $\mathcal{D}$  with the property that  $\Gamma(t, D') = \Gamma(t', D')$  for all  $t, t' \in D, D' \in \mathcal{D}$ . For simplicity assume that  $K$  is finite and hence each  $D \in \mathcal{D}$  is finite. Consider any two sets  $D_1, D_2 \in \mathcal{D}$  and note that, since  $\gamma$  is weakly increasing, the highest type in  $D_2$  must have the same maximal preference against types in  $D_1$  as the lowest type in  $D_2$  (first equality below). Similarly, the highest type in  $D_1$  must have the same minimal preference against types in  $D_2$  as the lowest type (second equality below). Hence,

$$\Gamma(\max D_1, \max D_2) = \Gamma(\max D_1, \min D_2) = \Gamma(\min D_1, \min D_2)$$

The two equalities and the monotonicity of  $\gamma$  imply that the types in  $D_1$  must behave identically against all opponent types - and hence  $D_1$  is a singleton set.

To see the converse, take any reciprocity model  $M = (T, \gamma')$ . A standard utility representation argument ensures that the types' kindness can be represented by a real-valued function. Consider two equally nice types, i.e., two types  $t, t'$  with  $\gamma(t, \cdot) = \gamma(t', \cdot)$ . Reciprocity implies that all other types must treat  $t$  and  $t'$  the same and then, validity implies that  $t = t'$ . Hence,  $\zeta_{\gamma'}$  is strictly increasing.

A simple example of a reciprocity model is the *linear reciprocity model* of Example 3 below:

**Example 3:** Let  $K = [0, 1]$  and

$$\gamma(x, y) = a + bx + cy + dxy$$

where  $b > 0, d > -b$ . The parameter  $d$  reveals the complementarity between reciprocity and kindness. If  $d > 0$ , then nicer types reciprocate more than less nice types whereas for  $d < 0$ , nicer types reciprocate less.<sup>10</sup> Levine's (1998) model of behavioral types corresponds to Example 3 with  $d = 0$ .

The reciprocity model can be used to address well-known experimental findings of the ultimatum bargaining game and related games. (See, for example, Blount (1995)). Experiments have shown that players are more likely to reject a given low offer in a standard ultimatum bargaining game than in the game in which offers are drawn at random by the experimenter. We can use the reciprocity model described above to interpret this regularity. Consider a setting in which players are uncertain about the opponent's personality. Then, if the opposing player makes a low offer, the responder can infer his opponent is not nice and, as a result, may reject the offer. When the randomization device makes the offer, no inference can be made about the opponent's type and, as a result, the responder is less inclined to reject a low offer.<sup>11</sup>

---

<sup>10</sup> This interpretation is appropriate provided that the real numbers identified with preferences can be interpreted as cardinal quantities; that is, provided that  $r_4 - r_3 = r_2 - r_1$  can be meaningfully interpreted as " $r_4$  is just as nicer than  $r_3$  as  $r_2$  is nicer than  $r_1$ ."

<sup>11</sup> For a worked out example illustrating this mechanism, see an earlier version of this paper. Gul and Pesendorfer (2005).

The ordered reciprocity model has the implication that types who make generous offers are least likely to reject offers while types who make the least generous offers are most likely to reject an offer. This is a consequence of the assumption that types are can be ranked according to how “nice” they are.<sup>12</sup>

## 5. Related Literature

Mertens and Zamir (1985) and Brandenburger and Dekel (1993) define a type as an infinite hierarchy of beliefs over a set of possible parameters. Those parameters – or payoff types (Battigalli and Siniscalchi (2003)) – are by assumption exogenous and therefore require no further explanation.<sup>13</sup> The interdependence of Harsanyi types arises from the interaction of the agents’ beliefs. Agent 1’s type influences agent 2’s payoff because 1 has information about a payoff relevant parameter. Interdependence in our setting is not related to a player’s information.<sup>14</sup>

In our model, a player’s personality specifies how he reacts to the characteristics and personalities of other players. We require that each type profile correspond to a unique preference profile. This requirement captures the idea that a player’s personality and characteristic are a complete description of the relevant data and therefore fully determine the resulting preference profile. In models of asymmetric information, there is no analogous requirement and players’ types need not resolve all uncertainty about preferences.

Mariotti, Meier and Piccione (2004) provide Mertens-Zamir and Brandenburger-Dekel type foundations for possibility structures<sup>15</sup> that identifies a type  $t$  with the combinations of parameter values and opponent types that  $t$  considers possible. Mariotti, Meier and Piccione (2004) prove that each type in a possibility structure can be identified with a certainty hierarchy. IPMs can be interpreted as a special class of possibility structures: ones in which every type of one player is possible given any type of the other and each type profile is associated with a unique preference profile.

---

<sup>12</sup> We thank Larry Samuelson for a related observation.

<sup>13</sup> Battigalli and Siniscalchi’s payoff types are akin our characteristics.

<sup>14</sup> In fact, in this paper, we ignore the problem of asymmetric information altogether.

<sup>15</sup> A possibility structure is an interactive belief models in which a player is either certain that a particular assertion is true, or is certain that it is false, or believes that it may be true or false (hence, possible) but assigns no probability to it.

That types can be uniquely identified through preference statements is a central property of our model. The same objective – that differences in types can be identified with differences in payoff relevant primitives – can be pursued in a framework with exogenously given payoff-types.<sup>16</sup> Bergemann and Morris (2007) establish that this question plays a central role in robust virtual implementation. Their model permits asymmetric information (i.e., players form conjectures over the preference types of their opponents) and they prove two results in which conditions similar to validity play a role.

To facilitate the comparisons, we consider finite, symmetric two-person IPMs with a single characteristic.<sup>17</sup> Each pair of types yields a pair of von Neumann-Morgenstern utilities on  $\Delta(X)$ . Bergemann and Morris make a mild genericity assumption on the IPM; they assume that given any belief over opponents types and actions, each type  $t$  is never indifferent over all outcomes in  $X$ . Let  $M = \{\gamma, T\}$  be an IPM satisfying these conditions and assume that the two agents are playing an arbitrary two-person game  $G = \{A_1, A_2, g\}$ , where  $g : A_1 \times A_2 \rightarrow \Delta(X)$ . Hence,  $A_i$  is the set of pure strategies of player  $i$  and  $g$  is the outcome function that relates pure strategy profiles to the space of lotteries over which the players preferences are defined.

For any game IPM  $M$  and game  $G$ , Bergemann and Morris define the rationalizable strategies as follows: in each round, players are allowed any conjecture over the opponents types and pure strategies that have not been eliminated for those types. For each type, the strategies that are never best responses against such conjectures are eliminated. The strategies that are not eliminated in any round are the rationalizable strategies of type  $t$ .

Bergemann and Morris call two types  $t, t'$  strategically equivalent if for every game  $G$ , the set of rationalizable actions of  $t$  and  $t'$  are identical. To relate their Proposition 5 to our analysis of communicability, we present the following stronger notion of validity:

**Definition:** *The IPM  $M = (T, \gamma, \omega)$  is strongly valid if the only decomposition  $\mathcal{D}$  of  $T$  such that*

$$(i) \ t, t' \in D \in \mathcal{D} \text{ implies } \omega(t) = \omega(t')$$

---

<sup>16</sup> For example, Ely and Peski (2006) point out that in standard models of incomplete information, two different types may have exactly the same hierarchy of beliefs. See also Dekel, Fudenberg and Morris (2006a), (2006b) for related work. In an earlier version of this paper (Gul and Pesendorfer (2005)), we show that an IPM is valid if and only if each type is uniquely identified through its certainty hierarchy.

<sup>17</sup> Bergemann and Morris allow for arbitrary finite  $n$ -person IPMs.

(ii)  $t' \in D^t \in \mathcal{D}$  implies  $\gamma(t, D) = \gamma(t', D)$  for all  $D \in \mathcal{D}$  is the finest decomposition.

Note that the only difference between validity and strong validity is that  $\gamma$  replaces  $\Gamma$  in the latter. Hence, strong validity would be the appropriate concept for Theorem 1 (or Theorem 3) if each player were restricted to making statements about his own preferences. We can now state Proposition 5 of Bergemann and Morris as follows:

**Proposition:** *Suppose  $M = (\gamma, T)$  satisfies the genericity condition above. Then, if  $M$  fails strong validity there exist at least two equivalent types in  $T$ .*

We can relate the result above to Theorem 3 as follows: if two types cannot distinguish themselves through any (truthful) preference statement, then they certainly cannot distinguish themselves through their strategic behavior. Of course, a type may have knowledge about preferences that is not strategically relevant, for example, he may know facts about his opponent's preferences that the opponent does not know. Hence, validity is not enough to rule out strategically equivalent types but the failure of validity ensures that there are strategically equivalent types.

While formally related, Theorem 3 and the proposition above have different objectives and interpretations. Theorem 3 asks if types in a particular IPM can be interpreted as a legitimate, non-circular description of individuals' attitudes toward each other. It identifies the following test: for an IPM to be valid, given any type profile  $(t, t')$ , there should be a sequence of statements (about preferences) that would enable both players to figure out their opponents' types. Hence, we interpret validity as a constraint on the modeler; IPMs that fail validity will have types that cannot be distinguished except through the arbitrary notational devices of the modeler.

In contrast, Bergemann and Morris have in mind situations in which types have clear meaning; that is, they view types as privately observed characteristics. For example, suppose there are two kinds of two-way radios ( $A$  and  $B$ ). Suppose also that the two players derive a benefit from their radio only if both have the same kind and invest a dollar to activate their radios. Hence, there are two outcomes, activate (1) and don't activate (0). When a type  $A$  confronts a type  $A$  or a type  $B$  confronts a type  $B$ , both prefer 1 to 0. Otherwise, both prefer 0 to 1. In this example, if we interpret  $A$  and  $B$  as

types rather than characteristics validity fails. The proposition above implies that both types will have exactly the same set of rationalizable strategies in every game. This does not mean that the model is in any sense ill-defined; being type  $A$  or  $B$  has a clear meaning in this model. However, as Bergemann and Morris show, no social choice rule that treats types  $A$  and  $B$  differently can be robustly implemented by a designer who does not know the players' types.<sup>18</sup>

Aumann (1976) shows that in a finite asymmetric information model with a common prior if the posteriors are common knowledge, then they must be identical. Geanakoplos and Polemarchakis (1982) investigate how posteriors might become common knowledge. They show that if two agents exchange information by sequentially revealing their current probability assessments (of a particular event), then, eventually, these assessments will become common knowledge (and hence, common if the priors are common as well). The subsequent literature on communication and consensus extends this result in the following ways: the function being communicated is not just priors but an arbitrary mapping from the set of all events,<sup>19</sup> there are more than two communicating agents and explicit, general protocols determining who speaks when.<sup>20</sup>

In these papers, the medium of communication; that is, the language is an arbitrary function  $g : 2^S \setminus \{\emptyset\} \rightarrow Y$  such that

$$g(E) = g(E'), \quad E \cap E' = \emptyset \text{ implies } g(E \cup E') = g(E) \quad (C)$$

Agents take turns announcing  $g(E)$  to some subset of other agents, where  $E \subset S$  is the smallest event that the agent knows to be true given all that he has heard before. These papers identify conditions on  $g$  and the protocol that ensure that eventually all agents have the same knowledge about the value of  $g$ . Despite the absence of priors in our model, some comparisons between Theorem 3 and the results in this literature are possible. Our language of preferences yields the following function  $g$ :

$$g(E) = \{\nu(s) \mid s \in E\}$$

---

<sup>18</sup> Bergemann and Morris' main theorem shows that a condition stronger than strong validity is necessary and sufficient to ensure that for any distinct  $t, t'$ , there exist some game  $G$  in which set of rationalizable strategies of  $t$  and  $t'$  are disjoint. They show that the latter property plays a key role in robust implementation with simultaneous mechanisms.

<sup>19</sup> See for example, Cave (1983) and Bacharach (1985).

<sup>20</sup> See Parikh and Krasucki (1990).

Thus, an agent who knows the event  $E$ , knows that the true preference profile is in  $g(E)$ . Such a  $g$  satisfies the *convexity* condition (C) above. The standard consensus result of the literature corresponds to the assertion that  $\mathcal{T}_1 * G(\mathcal{L}) = \mathcal{T}_2 * G(\mathcal{L})$ ; that is, once communication stops the two agents have the same knowledge about preferences (i.e., the function  $g$ ). Note that our focus is on whether types can be communicated in the language of preferences rather than whether communicating in the language of preferences eventually leads to agreement about preferences.

Our notion of communication is more permissive than the Caves-Bacharach-Parikh and Krasucki model of communication protocols. Our modeling has the effect of permitting conditional statements such as “had you told me  $x$ , I would have said  $y$ .” A communication protocol does not permit such statements. Hence, our version of communication always leads to (weakly) more “knowledge sharing” than any communication protocol. Furthermore, there are examples in which our model can convey knowledge that cannot be conveyed through all possible protocols.

## 6. Appendix

Let  $Z$  be a compact metric space. For any sequence  $A_n \in \mathcal{H}_Z$ , let

$$\underline{\lim} A_n = \{z \in Z \mid z = \lim z_n \text{ for some sequence } z_n \text{ such that } z_n \in A_n \text{ for all } n\}$$

$$\overline{\lim} A_n = \{z \in Z \mid z = \lim z_{n_j} \text{ for some sequence } z_{n_j} \text{ such that } z_{n_j} \in A_{n_j} \text{ for all } j\}$$

Let  $X$  be a metric space and  $p : X \rightarrow \mathcal{H}_Z$ . We say that  $p$  is Hausdorff continuous if it is a continuous mapping from the metric space  $X$  to the metric space  $\mathcal{H}_Z$ . Note that if  $p$  is a Hausdorff continuous mapping from  $X$  to  $\mathcal{H}_Z$ , then  $p \in \mathcal{C}(X, \mathcal{H}_Z)$ . However, the converse is not true.

**Lemma 1:** *Let  $X, Y, Y', Z$  be nonempty compact metric spaces,  $q \in \mathcal{C}(X \times Y, Z)$ ,  $p \in \mathcal{C}(Y', \mathcal{H}_Y)$ , and  $r \in \mathcal{C}(Y', Y)$ . Then, (i)  $A_n \in \mathcal{H}_Z$  converges to  $A$  (in the Hausdorff topology) if and only if  $\underline{\lim} A_n = \overline{\lim} A_n = A$ . (ii)  $x_n \in X$  converges to  $x$  implies  $q(x_n, B)$  converges to  $q(x, B)$  for all  $B \in \mathcal{H}_Y$ . (iii) If  $q^*(x, y') = q(x, p(y'))$  for all  $x \in X, y' \in Y'$  then  $q^* \in \mathcal{C}(X \times Y', \mathcal{H}_Z)$ . (iv) If  $r$  is onto, then  $r^{-1} \in \mathcal{C}(Y, \mathcal{H}_{Y'})$ .*

**Proof:** Part (i) is a standard result. See Brown and Percy (1995).

(ii) Suppose  $x_n \in X$  converges to  $x$ . Let  $z_{n_j} \in q(x_{n_j}, B)$  such that  $\lim z_{n_j} = z$ . Hence,  $z_{n_j} = q(x_{n_j}, y_{n_j})$  for some  $y_{n_j} \in B$ . Since  $B$  is compact, we can without loss of generality assume  $y_{n_j}$  converges to some  $y \in B$ . Hence, the continuity of  $q$  ensures  $z = q(x, y)$  and therefore  $z \in q(x, B)$  proving that  $\overline{\lim}q(x_n, B) \subset q(x, B)$ . If  $z \in q(x, B)$ , then there exists  $y \in B$  such that  $z = q(x, y)$ . Since  $q$  is continuous, we have  $z = \lim q(x_n, y)$ . Hence,  $q(x, B) \subset \underline{\lim}q(x_n, B)$ . Since,  $\underline{\lim}q(x_n, B) \subset \overline{\lim}q(x_n, B) \subset q(x, B)$ , we conclude  $\underline{\lim}q(x_n, B) = q(x_n, B) = \overline{\lim}q(x_n, B)$  as desired.

(iii) Suppose  $(x_n, y'_n)$  converges to  $(x, y)$  and  $z_n \in q^*(x_n, y'_n)$  converges to  $z$ . Pick  $y_n \in p(y'_n)$  such that  $q(x_n, y_n) = z_n$ . Since  $Y$  is compact, we can assume that  $y_n$  converges to some  $y$ . Since  $p \in \mathcal{C}(Y', \mathcal{H}_Y)$ , we conclude that  $y \in p(y')$  and since  $q$  is continuous,  $q(x, y) = z$ . Therefore,  $z \in q^*(x, y')$ , proving that  $q^* \in \mathcal{C}(X \times Y, \mathcal{H}_Z)$ .

(iv) The continuity and ontoness of  $r$  ensures that  $r^{-1}$  maps  $Y$  into  $h_{Y'}$ . Assume that  $y_n$  converges to  $y$ ,  $y'_n \in r^{-1}(y_n)$  and  $y'_n$  converges to  $y'$ . Then,  $r(y'_n) = y_n$  for all  $n$  and by continuity  $r(y') = y$ . Therefore,  $y' \in r^{-1}(y)$  as desired.  $\square$

**Lemma 2:** *Let  $X$  and  $Z$  be compact metric spaces. Suppose  $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$  and  $p_n(x) \subset p_{n+1}(x)$  for all  $n \geq 1$ ,  $x \in X$ . Let  $p(x) := \bigcap_{n \geq 1} p_n(x)$  and assume  $p(x)$  is a singleton for all  $x \in X$ . Then, (i)  $p$  is continuous and (ii)  $p_n$  converges to  $p$ .*

**Proof:** Obviously,  $\bigcap_{n \geq 1} G(p_n) = G(p)$ . Since  $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$  and  $X, Z$  are compact, so is  $G(p_n)$ . Therefore  $G(p)$  is compact (and therefore closed) as well. Since  $p$  is a function and both  $X, Z$  are compact, the fact that  $p$  has a closed graph implies that  $p$  is continuous.

To prove (ii), it is enough to show that if  $G_n$  is a sequence of compact sets such that  $G_{n+1} \subset G_n$  then  $G_n$  converges (in the Hausdorff topology) to  $G := \bigcap_n G_n$ . If not, since  $G_1$  is compact, we could find  $\epsilon > 0$  and  $y_n \in G_n$  converging to some  $y \in G_1$  such that  $d(y_n, G) > \epsilon$  for all  $n$ . Hence,  $d(y, G) \geq \epsilon$  and therefore there exists  $k$  such that  $y \notin G_k$  for all  $n \geq k$ . Choose  $\epsilon' > 0$  such that  $\min_{y' \in G_k} d(y', y) \geq \epsilon'$  and  $k'$  such that  $n \geq k'$  implies  $d(y_n, y) < \epsilon'/2$ . Then, for  $n \geq \max\{k, k'\}$  we have  $d(y_n, y) \geq \epsilon'$  and  $d(y_n, y) < \epsilon'/2$ , a contradiction.  $\square$

We say that  $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$  converges to  $p \in \mathcal{C}(X, \mathcal{H}_Z)$  uniformly if for all  $\epsilon > 0$ , there exists  $N$  such that  $n \geq N$  implies  $d(p_n(x), p(x)) < \epsilon$ .

Let  $X$  be an arbitrary set and  $Z$  be a compact metric space. Given any two functions  $p, q$  that map  $X$  into  $\mathcal{H}_Z$ , let  $d^*(p, q) = \sup_{x \in X} d(p(x), q(x))$ , where  $d$  is the Hausdorff metric on  $\mathcal{H}_Z$ .

**Lemma 3:** (i) If  $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$  converges to  $p \in \mathcal{C}(X, Z)$ , then  $p_n$  converges to  $p$  uniformly; that is,  $\lim_n d^*(p_n, p) = 0$ . (ii) The relative topology of  $\mathcal{C}(X, Z) \subset \mathcal{C}(X, \mathcal{H}_Z)$  is the topology of uniform convergence.

**Proof:** Let  $\lim p_n = p \in \mathcal{C}(X, Z)$ . Then,  $p$  is continuous and since  $X$  is compact, it is uniformly continuous. For  $\epsilon > 0$  choose a strictly positive  $\epsilon' < \epsilon$  such that  $d(x, x') < \epsilon'$  implies  $d(p(x), p(x')) < \epsilon$ . Then, choose  $N$  so that  $d_H(G(p), G(p_n)) < \epsilon'$  for all  $n \geq N$ . Hence, for  $n \geq N$ ,  $x \in X$  and  $z \in p_n(x)$ , we have  $x' \in X$  such that  $d(x, x') < \epsilon'$  and  $d(p(x'), z) < \epsilon'$ . Hence,  $d(p(x), z) \leq d(p(x'), z) + d(p(x'), p(x)) < 2\epsilon$  as desired.

Next, we will show that  $p_n$  converges to  $p$  uniformly implies  $G(p_n)$  converges to  $G(p)$  in the Hausdorff metric. This, together with (i) will imply (ii). Consider any sequence  $p_n$  converging uniformly to  $p$ . Choose  $N$  such that  $n \geq N$  implies  $d(p_n(x), p(x)) \leq \epsilon$ . Hence, for  $n \geq N$ ,  $(x, z) \in G(p_n)$  implies  $d((x, z), (x, p(x))) < \epsilon$ , proving  $\overline{\lim} G(p_n) \subset G(p) \subset \underline{\lim} G(p_n)$ .  $\square$

For  $\theta_n \in \Theta_n$  and  $n \geq 0$ , let

$$\Theta(\theta_n) = \{\theta' \in \Theta \mid \theta'(n) = \theta_n\}$$

**Lemma 4:** Let  $\hat{\theta} \in \Theta \in \mathcal{I}$  with  $\hat{\theta} = (f_0, f_1, \dots)$  and  $\phi(\hat{\theta}) = (f_0, f)$ . Then, for all  $n \geq 1$  and  $\theta_{n-1} \in \Theta_{n-1}$ ,  $\bigcup_{\theta \in \Theta(\theta_{n-1})} f(\theta) = f_n(\theta_{n-1})$ .

**Proof:** Let  $P \in f_n(\theta_{n-1})$ . Since the sequence  $\{\Theta_n\}$  is consistent, we may choose  $\theta_n \in \Theta_n(\theta_{n-1})$  so that  $P \in f_{n+1}(\theta_n)$ . Repeat the argument for every  $k > n$  to obtain  $\theta = (\theta_{n-1}, g_n, g_{n+1}, \dots) \in \Theta$  such that  $\phi(\hat{\theta})(\theta) = P$ . Hence,  $f_n(\theta) \subset \bigcup_{\theta \in \Theta(\theta_{n-1})} f(\theta) = f_n(\theta_{n-1})$ . That  $\bigcup_{\theta \in \Theta(\theta_{n-1})} f(\theta) \subset f_n(\theta_{n-1})$  follows from the definition of  $f$  and the fact that  $f_{n+1}(\theta) \subset f_n(\theta)$  for all  $n$  and all  $\theta \in \Theta$ .  $\square$

**Lemma 5:** Let  $X, Y$  be compact metric spaces and  $Z$  be an arbitrary metric space. Let  $q : X \times Y \rightarrow Z$  and let the mapping  $p$  from  $X$  to the set of functions from  $Y$  to  $Z$  be defined as  $p(x)(y) := q(x, y)$ . Then,  $q \in \mathcal{C}(X \times Y, Z)$  if and only if  $p \in \mathcal{C}(X, \mathcal{C}(Y, Z))$ .

**Proof:** Assume  $q$  is continuous. Since  $X \times Y$  is compact,  $q$  must be uniformly continuous. Hence, for all  $\epsilon > 0$  there exists  $\epsilon' > 0$  such that  $d((x, y), (x', y')) < \epsilon'$  implies  $d(q(x, y), q(x', y')) < \epsilon$ . In particular,  $d(x, x') < \epsilon'$  implies  $d(q(x, y), q(x', y)) < \epsilon$  for all  $y \in Y$ . Hence,  $d(x, x') < \epsilon'$  implies  $d(p(x), p(x')) < \epsilon$ , establishing the continuity of  $p$ . Next, assume that  $p$  is continuous and let  $\epsilon > 0$ . To prove that  $q$  is continuous, assume  $(x^k, y^k) \in X \times Y$  converges to some  $(x, y) \in X \times Y$ . The continuity of  $p$  ensures that for some  $k \in \mathbb{N}$ ,  $m \geq k$  implies  $d(p(x^m), p(x)) \leq \epsilon$ . Since  $p(x)$  is continuous, we can choose  $k$  so that  $d(p(x)(y^m), p(x)(y)) < \epsilon$  for all  $m \geq k$  as well. Hence,

$$d(p(x^m)(y^m), p(x)(y)) \leq d(p(x^m)(y^m), p(x)(y^m)) + d(p(x)(y^m), p(x)(y)) < 2\epsilon$$

□

**Lemma 6:** Let  $X$  be compact and  $Z$  be an arbitrary metric space. Suppose  $p \in \mathcal{C}(X, Z)$  is one-to-one. Then,  $p$  is a homeomorphism from  $X$  to  $p(X)$ .

**Proof:** It is enough to show that  $p^{-1} : p(X) \rightarrow X$  is continuous. Take any closed  $B \subset X$ . Since  $X$  is compact, so is  $B$ . Then,  $(p^{-1})^{-1}(B) = p(B)$  is compact (and therefore closed) since the continuous image of a compact set is compact. Hence, the inverse image of any closed set under  $p^{-1}$  is closed and therefore  $p^{-1}$  is continuous. □

**Lemma 7:** Let  $X, Y$  be a compact metric spaces. For  $p \in \mathcal{C}(X, \mathcal{H}_Y)$ , let  $\bar{d}(p) = \max_{x \in X} \max_{y, z \in p(x)} d(y, z)$ . Then, (i)  $p_n \in \mathcal{C}(X, \mathcal{H}_Y)$  converges to  $p \in \mathcal{C}(X, \mathcal{H}_Y)$  implies  $\limsup \bar{d}(p_n) \leq \bar{d}(p)$ . (ii)  $p, q, p', q' \in \mathcal{C}(X, \mathcal{H}_Y)$  and  $p(x) \subset p'(x), q(x) \subset q'(x)$  for all  $x \in X$  implies  $d(p, q) \leq \max\{d(p', q') + \bar{d}(p'), d(p', q') + \bar{d}(q')\}$ .

**Proof:** Since  $X \times Y$  is compact (i) is equivalent to the following:  $p_n \in \mathcal{C}(X, \mathcal{H}_Y)$  converges to  $p \in \mathcal{C}(X, \mathcal{H}_Y)$ ,  $\lim \bar{d}(p_n) = \alpha$  implies  $\alpha \leq \bar{d}(p)$ . To prove this, choose  $x_n \in X$  and  $y_n, z_n \in p_n(x_n)$  such that  $d(y_n, z_n) = \bar{p}_n$ . Without loss of generality, assume  $(x_n, y_n, z_n)$  converges to  $(x, y, z)$ . Since  $p_n$  converges to  $p$ , for all  $\epsilon > 0$ , there exists  $N$  such that

for all  $n \geq N$ , there exists  $(x'_n, y'_n)$  and  $(\hat{x}_n, \hat{z}_n)$  such that  $d((x'_n, y'_n), (x_n, y_n)) < \epsilon$  and  $d((\hat{x}_n, \hat{z}_n), (x_n, z_n)) < \epsilon$ . Hence, we can construct a subsequence  $n_j$  such that  $x'_{n_j}, \hat{x}_{n_j}$  both converge to  $x$ ,  $y'_{n_j}$  converges to  $y$ ,  $\hat{z}_{n_j}$  converges to  $z$ , and  $y'_{n_j} \in p(x'_{n_j}), \hat{z}_{n_j} \in p(\hat{x}_{n_j})$  for all  $n_j$ . Since  $p \in \mathcal{C}(X, \mathcal{H}_Y)$  we conclude  $y, z \in p(x)$ . But  $\alpha = \lim \bar{p}_n = \lim d(y_n, z_n) = d(y, z)$ . Hence,  $\alpha \leq \bar{d}(p)$ .

(ii) Let  $(x, z) \in G(p), (\hat{x}, \hat{z}) \in G(p')$ . Then,

$$d((x, z), (\hat{x}, \hat{z})) \leq \min_{(\hat{x}, \hat{y}) \in G(q')} d((x, z), (\hat{x}, \hat{y})) + \bar{d}(p')$$

Therefore,

$$\min_{(\hat{x}, \hat{z}) \in G(q)} d((x, z), (\hat{x}, \hat{z})) \leq d(p', q') + \bar{d}(q')$$

and a symmetric argument shows that

$$\min_{(x, z) \in G(p)} d((x, z), (\hat{x}, \hat{z})) \leq d(p', q') + \bar{d}(p')$$

Therefore,  $d(p, q) \leq \max\{d(p', q') + \bar{d}(p'), d(p', q') + \bar{d}(q')\}$ .  $\square$

### 6.1 Proof of Theorem 1:

We first show that  $\phi$  is continuous. Consider any sequence  $\theta^k = (f_0^k, f_1^k, \dots) \in \Theta$  such that  $\lim \theta^k = \theta = (f_0, f_1, \dots) \in \Theta$ . Let  $\phi(\theta) = (f_0, f)$  and  $\phi(\theta^k) = (f_0^k, f^k)$  for all  $k$ . Let  $\theta^k = (f_0^k, f_1^k, \dots), \theta = (f_0, f_1, \dots)$  and  $\epsilon > 0$ . By Lemma 2  $f_n$  converges to  $f$  and therefore by Lemma 7(i) there exists  $N$  such that  $\bar{d}(f_N) < \epsilon$ . Since  $f_N^k \rightarrow f_N$  Lemma 7(i) implies that there exists  $k'$  such that for  $k \geq k'$ ,  $\bar{d}(f_N^k) \leq 2\epsilon$ . Finally, there is  $k''$  such that  $d(f_N^k, f_N) \leq \epsilon$  for  $k > k''$ . Let  $m = \max\{k', k''\}$ . Lemma 7(ii) now implies that  $d(f_n^k, f_n) \leq 3\epsilon$ , for all  $n \geq N$  and  $k \geq m$ . Therefore  $d(f^k, f) \leq 3\epsilon$  for all  $k \geq m$ . This shows that  $\phi$  is continuous.

Next, we prove that  $\phi$  is one-to-one. Pick any  $(f_0, f_1, \dots), (g_0, g_1, \dots) \in \Theta$ . Let  $(f_0, f) = \phi(f_0, f_1, \dots)$  and  $(g_0, g) = \phi(g_0, \dots)$ . If  $f_0 \neq g_0$ , then clearly  $(f_0, f) \neq (g_0, g)$ . Hence, assume  $f_0 = g_0$ . Then, there exists a smallest  $n \geq 1$  and  $\theta_{n-1} \in \Theta_{n-1}$  such that  $g_n(\theta_{n-1}) \neq f_n(\theta_{n-1})$ . By Lemma 4,  $\bigcup_{\theta' \in \Theta(\theta_{n-1})} f(\theta') \neq \bigcup_{\theta' \in \Theta(\theta_{n-1})} g(\theta')$  and hence  $f \neq g$  as desired.

Since  $\phi$  is continuous and one-to-one and  $\Theta$  is compact, it follows from Lemma 6 that  $\phi$  is a homeomorphism from  $\Theta$  to  $\phi(\Theta)$ . The continuity of  $\Psi$  follows from the compactness of  $\Theta$  and Lemma 5.  $\square$

## 6.2 Proof of Theorem 2:

We say that  $\mathcal{D}$  is strongly continuous if the function  $\sigma : X \rightarrow \mathcal{D}$  defined by  $\sigma(x) = D^x$  is an element of  $\mathcal{C}(X, \mathcal{H}_X)$ .

Let  $M = (T, \gamma, \omega)$  be an IPM. Define the sequence of decompositions  $\mathcal{D}_n$  on  $T$  as follows:

$$D_0^t = \{t' \in T \mid \omega(t') = \omega(t)\}$$

and  $\mathcal{D}_0 = \{D_0^t \mid t \in T\}$ . For  $n \geq 1$  we define inductively

$$D_n^t := \{t' \in D_{n-1}^t \mid \Gamma(t', D) = \Gamma(t, D) \text{ for all } D \in \mathcal{D}_{n-1}\}$$

and  $\mathcal{D}_n = \{D_n^t \mid t \in T\}$ . Let  $\mathcal{D} = \left\{ \bigcap_n D_n^t \mid t \in T \right\}$  and note that  $\mathcal{D}$  is a decomposition of  $T$ .

**Step 1:** (i) Each  $\mathcal{D}_n$  is continuous. (ii)  $M$  is valid if and only if  $\mathcal{D} = \{\{t\} \mid t \in T\}$ .

**Proof:** (i) The proof is by induction. Assume that  $t_k$  converges to  $t$ ,  $\hat{t}_k \in D_0^{t_k}$  and  $\hat{t}_k$  converges to  $\hat{t}$ . Then,  $\omega(\hat{t}) = \lim \omega(\hat{t}_k) = \lim \omega(t_k) = \omega(t)$ . Hence,  $\hat{t} \in D_0^t$ , proving the strong continuity of  $\mathcal{D}_0$ . Assume that  $\mathcal{D}_n$  satisfies strong continuity. Hence, every  $D \in \mathcal{D}_n$  is compact. Assume that  $t_k$  converges to  $t$ ,  $\hat{t}_k \in D_{n+1}^{t_k}$  and  $\hat{t}_k$  converges to  $\hat{t}$ . Hence,  $\hat{t}_k \in D_n^{t_k}$  and by the strong continuity of  $\mathcal{D}_n$ , we have  $\hat{t} \in D_n^t$ . Pick any  $D \in \mathcal{D}_n$  and  $P \in \Gamma(\hat{t}, D)$ . By Lemma 1(ii), we have  $P_n \in \Gamma(\hat{t}_n, D) = \Gamma(t_n, D)$  such that  $\lim P_n = P$ . Then, by Lemma 1(iii), we have  $P \in \Gamma(t, D)$ , proving that  $\Gamma(\hat{t}, D) \subset \Gamma(t, D)$ . A symmetric argument ensured that  $\Gamma(\hat{t}, D) = \Gamma(t, D)$ , establishing that  $\hat{t} \in D_{n+1}^t$  and proving the strong continuity of  $\mathcal{D}_{n+1}$ . This concludes the proof of part (i).

If  $\mathcal{D} \neq \{\{t\} \mid t \in T\}$ , then  $\mathcal{D}$  is a challenging decomposition. Therefore,  $M$  is not valid. Suppose  $M$  is not valid and hence there exists a continuous decomposition  $\mathcal{D}^*$  that challenges  $M$ . Then,  $\mathcal{D}^*$  is a refinement of  $\mathcal{D}$ ; that is,  $D_t^* \in \mathcal{D}^*$  and  $D_t \in \mathcal{D}_t$  implies  $D_t^* \subset D_t$ . To see this note that since  $\mathcal{D}^*$  challenges  $M$  it is a refinement of  $\mathcal{D}^0$ . Moreover,

if  $\mathcal{D}^*$  is a refinement of  $\mathcal{D}^k$  then  $\mathcal{D}^*$  is a refinement of  $\mathcal{D}^{k+1}$ . Then last assertion follows from the fact that for  $t' \in D_t^* \in \mathcal{D}^*$ ,

$$\Gamma(t, D^k) = \bigcup_{D \in \mathcal{D}^*, D \subset D^k} \Gamma(t, D) = \bigcup_{D \in \mathcal{D}^*, D \subset D^k} \Gamma(t', D) = \Gamma(t', D^k)$$

Hence,  $\mathcal{D}^*$  is a refinement of  $\mathcal{D}^k$  for all  $k \geq 0$ . Hence,  $D_t^* \in \mathcal{D}^*$  implies

$$D_t^* \subset \left\{ \bigcap_k D_t^k \mid t \in T \right\} = D_t \in \mathcal{D}$$

This concludes the proof of step 1. □

Let  $\Theta_0 := \Omega = \omega(T)$ . Define  $f_0^t := \omega(t)$  and  $\iota_0(t) := f_0^t$  for all  $t \in T$  and define inductively  $f_n^t : \Theta_{n-1} \rightarrow \mathcal{H}$ ,  $\Theta_n, \iota_n : T \rightarrow \Theta_n$  as follows:

$$\begin{aligned} f_n^t(\theta_{n-1}) &= \Gamma(t, \iota_{n-1}^{-1}(\theta_{n-1})) \\ \iota_n(t) &= (\iota_{n-1}(t), f_n^t) \\ \Theta_n &= \iota_n(T) \end{aligned}$$

Let

$$\Theta = \{(f_0, f_1, \dots) \mid (f_0, f_1, \dots, f_n) \in \Theta_n \text{ for all } n \geq 0\}$$

$$\iota(t) = (f_0, f_1, \dots) \text{ such that } (f_0, f_1, \dots, f_n) = \iota_n(t) \text{ for all } n.$$

Henceforth, for any  $t \in T$  such that  $\iota(t) = (f_0, f_1, \dots)$  we write  $f_n^t$  to denote the corresponding  $f_n$ . We also define the functions  $g_n^t : T \rightarrow \mathcal{H}$  as

$$g_n^t(s) = \Gamma(t, D_{n-1}^s)$$

**Fact 1:** *For all  $n$ , the functions  $\iota_n$  are onto and continuous and the sets  $\Theta_n$  are non-empty and compact.*

**Proof:** We will prove inductively that  $\Theta_n$  are nonempty, compact,  $\iota_n$  is continuous and onto for every  $n$ . Clearly, this statement is true for  $n = 0$ . Suppose it is true for  $n$ . Then, by Lemma 1 parts (iii) and (iv),  $\iota_{n+1} \in \mathcal{C}(\Theta_n, \mathcal{H})$  and  $\Theta_{n+1}$  is compact. The functions  $\iota_n$  is onto by definition. □

**Fact 2:** (i) The function  $\iota$  is onto. (ii)  $\iota_n(t) = \iota_n(s)$  if and only if  $D_n^t = D_n^s$ . (iii)  $f_n^t(\iota_{n-1}(s)) = g_n^t(s)$ .

**Proof:** Next, we show that  $\iota : T \rightarrow \Theta$  is onto. Pick  $(f_0, f_1, \dots)$  such that  $(f_0, f_1, \dots, f_n) \in \Theta_n$  for all  $n$ . Then, for all  $n$ , there exists  $t_n \in T$  such that  $\iota_n(t_n) = (f_0, f_1, \dots, f_n)$ . Take  $t_{n_j}$ , a convergent subsequence of  $t_n$  converging to some  $t \in T$ . For all  $n$  and  $n_j > n$ ,  $\iota_n(t_{n_j}) = (f_0, f_1, \dots, f_n)$ . Hence, the continuity of  $\iota_n$  ensures that  $\iota_n(t) = (f_0, f_1, \dots, f_n)$  for all  $n$ , establishing that  $\iota(T) = \Theta$ .

Next, we prove that  $\iota_n(t) = \iota_n(s)$  if and only if  $D_n^t = D_n^s$ . To see this, note that for  $n = 0$ , the assertion is true by definition. Suppose, it is true for  $n$ . Then, if  $s \in D_{n+1}^t$ , we have  $s \in D_n^t$  and  $\Gamma(t, D_n) = \Gamma(s, D_n)$  for all  $D \in \mathcal{D}_n$ . Hence,  $f_{n+1}^t = f_{n+1}^s$  and therefore, by the inductive hypothesis,  $\iota_{n+1}(t) = \iota_{n+1}(s)$ . Conversely, if  $\iota_{n+1}(t) = \iota_{n+1}(s)$ , then  $f_{n+1}^t = f_{n+1}^s$  and  $i_n^t = i_n^s$ . Therefore, by the inductive hypothesis,  $s \in D_{n+1}^t \in \mathcal{D}_{n+1}$ .

Part (iii) follows from part (ii) and the definitions of  $g_n^t, f_n^t$ .  $\square$

**Fact 3:** If  $M$  is valid then (i)  $g^t = \lim g_n^t$  is well defined and continuous and (ii)  $d(g_n^{t_n}, g) \rightarrow 0$  if  $t_n \rightarrow t$  as  $n \rightarrow \infty$ .

**Proof:** Part (i) follows from Lemma 2. For part (ii) fix  $\epsilon > 0$  and note that by Lemmas 3(i), 7(i) there exists  $N$  such that  $d(g^t, g_N^t) < \epsilon$  and  $\bar{d}(g_N^t) < \epsilon$ . By Lemma 1(ii)  $g_N^{t_n} \rightarrow g_N^t$ . By Lemma 7(i) we can choose  $m$  so that  $\bar{d}(g_N^{t_k}) \leq 2\epsilon$  for all  $n \geq m$ . Therefore, by Lemma 7(ii),  $d(g_n^{t_n}, g_n) < 3\epsilon$  for all  $n > \max\{m, N\}$ . It follows that  $d(g_n^{t_n}, g) < 4\epsilon$  for all  $n > \max\{m, N\}$  as desired.  $\square$

**Step 2:**  $M$  is isomorphic to some  $\Theta \in \mathcal{I}$  if and only if  $\mathcal{D} = \{\{t\} \mid t \in T\}$ .

Fact 2(ii) implies that  $\Theta_{n-1}(\theta_{n-2}) = \iota_{n-1}(D_{n-2}^s)$  for  $s$  such that  $\iota_{n-2}(t) = D_{n-2}^s$ . Therefore,

$$f_n^t(\theta_{n-2}) = \Gamma(t, D_{n-2}^s) = \bigcup_{s' \in D_{n-2}^s} \Gamma(t, D_{n-1}^{s'}) = \bigcup_{\theta'_{n-1} \in \Theta_{n-1}(\theta_{n-2})} f^t(\theta'_{n-1})$$

proving that  $\{\Theta_n\}$  satisfies the consistency condition.

Let  $f^t : \Theta \rightarrow \mathcal{H}$  be defined by

$$f^t(\theta) = \bigcap_{n \geq 1} f_n^t(\theta)$$

Assume that  $M$  is valid and hence  $\mathcal{D} = \{\{t\} \mid t \in T\}$ . Since,  $\iota_n(t) = \iota_n(s)$  if and only if  $D_n^t = D_n^s$  (Fact 2(ii)), we conclude that  $\iota$  is one-to-one. For  $\theta = (g_0, g_1, \dots)$  we let  $\theta(n)$  be defined as  $(g_0, \dots, g_n)$ . By Fact 2,  $f_n^t(\theta(n-1)) = g_n^t(s) = \Gamma(t, D_n^s)$  for  $\theta(n) = \iota(s)$ . It follows that  $f^t(\theta) = \Gamma(t, s)$  and therefore  $f^t$  is a singleton.

To prove that  $\iota$  is a homeomorphism, we prove that  $\iota$  is continuous and appeal to Lemma 6. Consider  $t_k$  converging to  $t$ . It follows from Fact 3 that for any two subsequences of natural numbers  $n(j), k(j)$  both converging to  $\infty$ ,  $g_{n(j)}^{t_{k(j)}}$  converges to  $g^t$ . Recall that  $d^*$  is the sup metric. It follows from Lemma 3(i) that  $g_{n(j)}^{t_{k(j)}}$  converges to  $g^t$  in the sup metric  $d^*$  as well. Hence, for any  $\epsilon > 0$ , there exists  $N$  such that  $k \geq N, n \geq N, d^*(g_n^{t_k}, g) < \epsilon$ . Since each  $\iota_n$  is continuous, we can choose  $k > N$  large enough so that  $d(f_n^k, f_n) < \epsilon$  for all  $n \leq N$ . Hence,

$$d(f_n^k, f_n) \leq d^*(f_n^k, f_n) = d^*(g_n^k, g_n) \leq d^*(g_n^k, g) + d^*(g, g_n) \leq 2\epsilon$$

proving the continuity of  $\iota$ . Note that

$$\psi(\iota(t), \iota(s)) = \bigcap_{n \geq 1} f_n^t(\iota(s)) = \lim \Gamma(t, D_n^s) = \Gamma(t, s)$$

Hence  $\Theta$  is isomorphic to  $M$  as desired.

Next we will show that if  $M$  is isomorphic to some  $\Theta$ , then the function  $\iota$  defined above is the isomorphism. Let  $\hat{\iota} : T \rightarrow \Theta$  be an isomorphism and  $\hat{\iota}_n$  denote the  $n$ -th coordinate function of  $\hat{\iota}$ . Recall that  $\iota$  defined above satisfies the property

$$\iota_n(t) = \iota_n(s) \text{ if and only if } D_n^t = D_n^s \tag{*}$$

Note that this property uniquely identifies the function  $\iota$ . That is, if  $\hat{\iota}$  is any function that also satisfies (\*),  $\hat{\iota} = \iota$ . To see this note that if  $\hat{\iota}_0$  satisfies (\*) then obviously,  $\hat{\iota}_0 = \omega = \iota_0$ . Then, a simple inductive step yields the desired conclusion. To see that  $\hat{\iota}$  satisfies (\*), note

that since it is an isomorphism, we have  $\omega = \hat{\iota}_0$  and hence  $(*)$  is satisfied for  $n = 0$ . Suppose it is satisfied for  $n$ . Then, suppose  $\hat{\iota}_{n+1}(t) = \hat{\iota}_{n+1}$ . Since  $\hat{\iota}$  is an isomorphism, we conclude  $f_{n+1}^t = f_{n+1}^s$ . Then, the inductive hypothesis yields  $D_{n+1}^t = D_{n+1}^s$ . Conversely, suppose  $D_{n+1}^t = D_{n+1}^s$ . Then,  $\Gamma(t, D_n) = \Gamma(s, D_n)$  for all  $D_n \in \mathcal{D}_n$ . Since,  $\hat{\iota}$  is an isomorphism, we conclude  $\psi(\hat{\iota}(t), \hat{\iota}(D_n)) = \psi(\hat{\iota}(s), \hat{\iota}(D_n))$  for all  $D_n \in \mathcal{D}_n$ . Which, by the inductive hypothesis, yields  $\hat{\iota}_{n+1}(t) = \hat{\iota}_{n+1}(s)$ .

Suppose  $s \in D_n^t \in \mathcal{D}_n$  for all  $n$ . Since  $\iota$  is an isomorphism, we have

$$f_n^t(\iota(D_n)) = \psi(\iota(t), \iota(D_n)) = \Gamma(t, D_n) = \Gamma(s, D_n) = \psi(\iota(s), \iota(D_n)) = f_n^s(\iota(D_n))$$

for all  $n, D_n \in \mathcal{D}_n$ . By  $(*)$ , we have  $\iota(t) = \iota(s)$ . Since  $\iota$  is one-to-one, we conclude  $s = t$ . This concludes the proof of step 2.  $\square$

Theorem 1 and Step 1 imply that any component of the canonical types space is a valid IPM. Steps 1 and 2 imply that any valid IPM is isomorphic to a component of the canonical type space.  $\square$

### 6.3 Proof of Theorem 3

Algebras and decompositions can be represented as functions: Let  $h : S \rightarrow X$  be any onto function. This  $h$  yields the decomposition  $\mathcal{T}_h = \{h^{-1}(x) \mid x \in X\}$  and the algebra  $\mathcal{A}_h := \{h^{-1}(V) \mid V \subset X\}$ . Conversely, for any decomposition  $\mathcal{T}$ , the canonical mapping  $\tau$  of  $\mathcal{T}$  (that is,  $h : S \rightarrow \mathcal{T}$  such that  $h(s)$  is the unique element of  $\mathcal{T}$  that contains  $s$ ) represents  $\mathcal{T}$  in this sense:  $\mathcal{T}_\tau = \mathcal{T}$ . When there is no risk of confusion, we will use a decomposition of  $\mathcal{T}$  and its canonical mapping  $\tau$  interchangeably. Note also that for any algebra  $\mathcal{A}$ ,  $\tau_{\mathcal{A}}$ , the set of minimal elements in  $\mathcal{A} \setminus \{\emptyset\}$  is a decomposition and  $\tau_{\mathcal{A}}$  interpreted as the canonical mapping represents  $\mathcal{A}$ ; that is,  $\mathcal{A} = \mathcal{A}_{\tau_{\mathcal{A}}}$ .

Let  $h : S \rightarrow X$  and  $k : S \rightarrow Y$  be two onto functions and define  $(h, k)(s) = (h(s), k(s))$  for all  $s$  and  $h * k : S \rightarrow Z$  for  $Z \subset 2^Y$  be the onto function defined by  $h * k(s) = k(h^{-1}(h(s)))$ .

**Fact:** For any two onto function  $h, k$  on  $S$ ,  $\mathcal{A}_h \vee \mathcal{A}_k = \mathcal{A}_{(h, k)}$  and  $\mathcal{A}_{h * k} = \mathcal{T}_h * \mathcal{A}_k$ .

**Proof:** The proof of the first assertion is straightforward. To prove the second assertion, we will show that (i)  $\mathcal{A}_{h * k}$  contains  $\mathcal{T}_h * A$  for all  $A \in \mathcal{A}_k$  and therefore, it contains

$\mathcal{T}_h * \mathcal{A}_k$  and (ii) for every  $x \in (h * k)(S)$ , there exist  $A^1, \dots, A^n \in \mathcal{T}_h * \mathcal{A}_k$  such that  $\bigcap_{m=1}^n A^m = (h * k)^{-1}(x)$ . (Hence,  $\mathcal{T}_h * \mathcal{A}_k$  contains  $\mathcal{T}_{h*k}$  and therefore it contains  $\mathcal{A}_{h*k}$ .)

For (i), let  $B = \mathcal{T}_h * A$ . Hence, there exist  $V \subset h(S)$  and  $W \subset k(S)$  such that (a) for all  $x \in V$ ,  $h(s) = x$  implies  $k(s) \in W$  and (b) for all  $x \notin V$  there exist  $s'$  such that  $h(s') = x$  and  $k(s') \notin W$ , and (c)  $B = h^{-1}(V)$ . To establish that  $B \in \mathcal{A}_{h*k}$ , we will show that  $s \in B$  and  $(h * k)(s) = (h * k)(\hat{s})$  implies  $\hat{s} \in B$ . Suppose  $s \in B$  and  $\hat{s} \notin B$ . Then, by (a) above  $(h * k)(s) \in W$  and by (b) there exists  $s'$  such that  $h(s') = h(\hat{s})$  and  $k(s') \notin W$ . Hence,  $(h * k)(s) \neq (h * k)(\hat{s})$  as desired.

To prove (ii), suppose  $x \in (h * k)(S)$ . Hence,  $x \subset k(S)$  and  $x \neq \emptyset$ . Let  $B = (h * k)^{-1}(x)$ . Let  $\{x_2, \dots, x_n\}$  be an enumeration of the set of all nonempty subsets of  $x$ . Define

$$A^1 = \{s \in S \mid k(\hat{s}) \in x \ \forall \hat{s} \in h(s)\}$$

$$A^m = \{s \in S \mid \text{there exists } \hat{s} \in h(s) \text{ such that } k(\hat{s}) \notin x\}$$

for  $m = 2, \dots, n$ . Note that  $A^1 = \mathcal{T} * k^{-1}(x)$ ,  $A^m = \mathcal{T} * (S \setminus k^{-1}(x_m))$  for  $m > 1$ , and  $B = \bigcap_{m=1}^n A^m$  as desired.  $\square$

For any IPM  $(T, \gamma)$ , we can construct an equivalent IPM-EF  $E = \{S, \mathcal{T}_1, \mathcal{T}_2, \nu\}$  as follows:  $S = T \times T$ ,  $\mathcal{T}_1 = \{\{t\} \times T \mid t \in T\}$ ,  $\mathcal{T}_2 = \{T \times \{t\} \mid t \in T\}$ , and  $\nu(t_1, t_2) = (\gamma(t_1, t_2), \gamma(t_2, t_1))$ . Let  $\zeta_1(t) = \{t\} \times T$ ,  $\zeta_2(t) = T \times \{t\}$ . Hence,  $E$  is equivalent to  $(T, \gamma)$ .

An algebra  $\mathcal{A}$  on  $S$  is symmetric if  $A \in \mathcal{A}$  implies  $\{(t_2, t_1) \in S \mid (t_1, t_2) \in A\} \in \mathcal{A}$ . Let  $\Lambda$  be the lattice of all symmetric algebras on  $S$ . The set  $\Lambda_{\mathcal{L}^p} := \{\hat{\mathcal{L}} \in \Lambda \mid \mathcal{L}^p \subset \hat{\mathcal{L}}\}$  is a sublattice of  $\Lambda$  and  $F$  is an increasing function on  $\Lambda_{\mathcal{L}^p}$ ; that is,

$$\mathcal{L}' \subset \mathcal{L}'' \text{ implies } F(\mathcal{L}') \subset F(\mathcal{L}'')$$

For  $\mathcal{L} \in \Lambda$ , the algebra  $G(\mathcal{L}) \in \Lambda_{\mathcal{L}}$  is a fixed-point of  $F$ . Let  $\hat{\mathcal{L}} \in \Lambda_{\mathcal{L}^p}$  be a fixed-point of  $F$ . Hence,  $\mathcal{L}^1 \subset \hat{\mathcal{L}}$  and therefore  $\mathcal{L}^2 = F(\mathcal{L}^1) \subset F(\hat{\mathcal{L}}) = \hat{\mathcal{L}}$ . By induction,  $G(\mathcal{L}) \subset F(\hat{\mathcal{L}}) = \hat{\mathcal{L}}$ . Hence,  $G(\mathcal{L})$  is the smallest fixed-point of  $F$  in  $\Lambda_{\mathcal{L}^p}$ .

Let  $\mathcal{A}^* \in \Lambda$  denote the richest algebra on  $S$ ; i.e.,  $\{s\} \in \mathcal{A}^*$  for all  $s \in S$ . Obviously,  $\mathcal{A}^*$  is a fixed-point of  $F$  and is the largest fixed point. If  $\mathcal{T}_1 \cup \mathcal{T}_2 \subset \mathcal{C}_{\mathcal{L}^p}$ , then  $\mathcal{A}^* = \mathcal{A}(\mathcal{T}_1 \cup \mathcal{T}_2) \subset \mathcal{C}_{\mathcal{L}^p} \subset G(\mathcal{L}^p)$ . Hence,  $\mathcal{A}^*$  is both the largest and smallest fixed-point of  $F$  in  $\Lambda_{\mathcal{L}^p}$ . So, if  $\mathcal{T}_1 \cup \mathcal{T}_2$  can be communicated in  $\mathcal{L}^p$ , then  $\mathcal{A}^*$  is the only fixed-point of  $F$  in  $\Lambda_{\mathcal{L}^p}$ .

Conversely, if  $\mathcal{A}^*$  is the only fixed-point in  $\Lambda_{\mathcal{L}^p}$ , then  $G(\mathcal{L}^p) = \mathcal{A}^*$ . Since,  $\mathcal{T}_i * \mathcal{A}^* = \mathcal{T}_i$ , we have  $\mathcal{C}_{\mathcal{L}^p} = \mathcal{T}_1 \vee \mathcal{T}_2 = \mathcal{A}^*$  and therefore  $\mathcal{T}_1 \cup \mathcal{T}_2 \subset \mathcal{C}_{\mathcal{L}^p}$ . Hence, the proposition is equivalent to the statement that  $\mathcal{A}^*$  is the only fixed-point of  $F$  in  $\Lambda_{\mathcal{L}^p}$ .

Suppose  $\mathcal{L} \neq \mathcal{A}^*$  is a fixed-point of  $F$  in  $\Lambda_{\mathcal{L}^p}$ . Then,  $\mathcal{T}_i * \mathcal{L} \neq \mathcal{A}(\mathcal{T}_i)$ . Let  $h_1, h_2, k$  be functions such that  $\mathcal{T}_{h_1} = \mathcal{T}_1$  and  $\mathcal{A}_k = \mathcal{L}$ . Consider the decomposition  $\mathcal{D}$  induced on  $T$  by the function  $h_1 \cdot \zeta_1$ . By symmetry, this is the same decomposition as the one induced by  $h_2 \cdot \zeta_2$ . Since  $\mathcal{T}_i * \mathcal{L} \neq \mathcal{A}(\mathcal{T}_i)$ ,  $\mathcal{D} \neq \{\{t\} \mid t \in T\}$ . Since  $\mathcal{L}$  is a fixed point of  $F$ , we can assume

$$k = (\nu, h_1 * k, h_2 * k) \quad (1)$$

Choose  $t' \in D^t$ , the element of  $\mathcal{D}$  that contains  $t$ . Hence,

$$k(h_2^{-1}(h_2(t, t^*))) = k(h_2^{-1}(h_2(t', t^*))) \quad (2)$$

It follows from equation (1) above that  $D^{t^*} \neq D^{t'}$  implies  $k(t, t^*) \neq k(t', t')$ . Therefore, equation (2) implies

$$k(\{t\} \times D^{t^*}) = k(\{t'\} \times D^{t'}) \quad (3)$$

Proving that for all  $\mathcal{T}^*$ , there exists  $\bar{t} \in D^{t^*}$  such that  $k(t', \bar{t}) = k(t, t^*)$  and therefore  $\nu(t', \bar{t}) = \nu(t, t^*)$ . Hence,  $\mathcal{D}$  challenges  $(T, \gamma)$ .

For the converse, let  $\mathcal{D}$  be a decomposition that challenges  $(T, \gamma)$ . Define  $h_1(t, t^*) = \{(t', t'') \mid t' \in D^t\}$  and  $h_2(t^*, t) = \{(t'', t') \mid t' \in D^t\}$ . It is easy to verify that  $\mathcal{A}_{(\nu, h_1, h_2)} \neq \mathcal{A}^*$  and  $\mathcal{A}_{(\nu, h_1, h_2)} \in \Lambda_{\mathcal{L}^p}$  is a fixed-point of  $F$ .  $\square$

#### 6.4 Proof of Theorem 4

Verifying the validity of  $M_k^0$  and  $M_k^1$  is straightforward for any  $k$ . To prove the converse, let  $(T', \gamma')$  be a valid binary symmetric model with  $\gamma'(x, y) \in \{0, 1\}$  for all  $x, y \in T'$  and consider the mapping  $\zeta_{\gamma'} : T' \rightarrow \mathcal{C}(T', \{0, 1\})$  defined by  $\zeta_{\gamma'}(x)(y) = \gamma'(x, y)$ . Since  $\gamma'$  is continuous, so is  $\zeta_{\gamma'}$ . Hence,  $\zeta_{\gamma'}(T')$  is compact and therefore finite.

Then, without loss of generality, assume  $T' = \{1, \dots, m\}$ . We will prove the result by induction. If  $k = 1$ , the result is obvious. Suppose that the result is true for  $k = m$  and assume  $m = k + 1$  and hence  $T' = \{1, \dots, k + 1\}$ . Validity ensures that  $\gamma'(t, \cdot)$  is constant

for some  $t \in T'$ . Suppose this constant is 1 and  $t = k + 1$ . (The proof for the case in which this constant is 0 is symmetric and will be omitted.) Let  $T = T' \setminus \{t\}$  and let  $\gamma$  be the restriction of  $\gamma'$  to  $T \times T$ .

First, we argue that  $(T, \gamma)$  must be a valid binary symmetric IPM. If not, there exist a challenging decomposition  $\mathcal{D}$ . Then, let  $\mathcal{D}' = \mathcal{D} \cup \{\{T\}\}$  and note that  $\mathcal{D}'$  challenges  $(T', \gamma')$ , contradicting the validity of  $(T', \gamma')$ . Hence, by the inductive hypothesis, we can assume that  $T = \{1, \dots, k\}$  and either  $\gamma = \gamma_k$  or  $\gamma = \gamma_{k+1}$ . In the former case, we have  $\gamma'(k, t') = \gamma(k+1, t') = 1$  for all  $t' \in T'$ , contradicting the validity of  $(T', \gamma')$ . In the latter case, we have  $(T', \gamma') = M_{k+1}^0$  as desired.  $\square$

## 6.5 Proof of Theorem 5:

**Lemma 9:** *Let  $K$  be any compact subset of the reals. Let  $\gamma : K \times K \rightarrow \mathbb{R}$  be weakly increasing in both arguments and continuous. Then,  $(K, \gamma)$  is a valid IPM if and only if  $\zeta_\gamma$  is strictly increasing.*

**Proof:** Suppose all the assumptions of the lemma are satisfied and  $\zeta_\gamma$  is strictly increasing. Hence, there exists  $x \neq z$  such that  $\gamma(x, y) = \gamma(z, y)$  for all  $y \in K$ . Then, define the decomposition  $\mathcal{D}$  as follows: for  $y \notin \{x, z\}$ ,  $D^y = \{y\}$ , and  $D^x = D^z = \{x, z\}$ . It follows that  $\gamma(x, \cdot) = \gamma(z, \cdot)$  and therefore  $\gamma(\cdot, x) = \gamma(\cdot, z)$  and hence  $\Gamma(w, D) = \Gamma(w', D)$  for all  $w \in K$ ,  $w' \in D^w$ , and  $D \in \mathcal{D}$ . Hence,  $(K, \gamma)$  is not valid.

Next, suppose that  $(K, \gamma)$  is not valid. Then, there exists a decomposition  $\mathcal{D}$  of  $K$  such that (i) there is  $D \in \mathcal{D}$  and  $x, z \in D$  such that  $x \neq z$ , (ii)  $\Gamma(w, D) = \Gamma(w', D)$  for all  $w \in K$ ,  $w' \in D^w$ , and  $D \in \mathcal{D}$ . Let  $\bar{D}$  denote the closure of  $D \in \mathcal{D}$ . The continuity of  $\Gamma$  ensures that

$$\Gamma(w, \bar{D}_2) = \Gamma(w', \bar{D}_2) \quad (*)$$

for all  $w, w' \in \bar{D}_1$  and  $D_1, D_2 \in \mathcal{D}$ . To see this, take  $w, w' \in \bar{D}_1$  and  $y \in \bar{D}_2$ . By definition, there exists a sequence  $(w_n, w'_n, y_n) \in D_1 \times D_1 \times D_2$  converging to  $(w, w', y)$ . Moreover, there exists  $y'_n \in D_2$  such that  $\Gamma(w_n, y_n) = \Gamma(w'_n, y'_n)$  for all  $n$ . Since  $\bar{D}_2$  is compact,  $y'_n$  has a convergent subsequence that converges to some  $y' \in \bar{D}_2$ . Assume, without loss of generality, that this subsequence is  $y'_n$  itself. Then, the continuity of  $\Gamma$  ensures  $\Gamma(w, y) = \Gamma(w', y')$  as desired.

The weak monotonicity of  $\gamma$  in both arguments together with (\*) implies

$$\Gamma(\max \bar{D}_1, \max \bar{D}_2) = \Gamma(\min \bar{D}_1, \max \bar{D}_2) = \Gamma(\min \bar{D}_1, \min \bar{D}_2)$$

Then, monotonicity of  $\Gamma$  ensures  $\gamma(w, y) = \gamma(w', y)$  for all  $y \in K$  whenever  $w, w' \in \bar{D}$ , in particular, for  $w = x$  and  $w' = z$ .  $\square$

Lemma 9 establishes that any  $M = (K, \gamma)$  such that  $\zeta_\gamma$  is strictly increasing is a reciprocity model. Next, suppose  $M = (T, \gamma)$  is a reciprocity model. Hence,  $\succeq$ , the nicer than relation is a preference relation. The continuity of  $\gamma$  yields the continuity of  $\succeq$ . Since  $T$  is a compact metric space, it is separable and hence there exists a continuous real-valued function  $x : T \rightarrow \mathbb{R}$  that represents  $\succeq$ . Let  $K := x(T) = \{x(t) \mid t \in T\}$ . Let  $D^t = \{t' \in T \mid x(t') = x(t)\}$  and  $\mathcal{D} = \{D^t \mid t \in T\}$ . Clearly,  $\mathcal{D}$  is a decomposition of  $T$  such that  $\gamma(t', \cdot) = \gamma(t, \cdot)$  for all  $t' \in D^t$ . Since every type reciprocates,  $\gamma(\cdot, t') = \gamma(\cdot, t)$  for all  $t' \in D^t$  and therefore  $\Gamma(t, D) = \Gamma(t', D)$  for all  $t \in T$ ,  $t' \in D^t$  and  $D \in \mathcal{D}$ . Since  $M$  is valid, each  $D^t$  is a singleton and therefore  $x$  is one-to-one. Then, the compactness of  $T$  ensures that  $K$  is compact and that  $x$  is a homeomorphism. Define,  $\gamma(w, y) = \gamma(x^{-1}(w), x^{-1}(y))$  for all  $w, y \in K$ . Since  $x$  and  $\gamma$  are continuous, so is  $\gamma'$ . It follows that  $(K, \gamma')$  is isomorphic to  $M$  and therefore is valid. Since  $x$  represents  $\succeq$  and every type in  $M$  reciprocates,  $\gamma'$  is weakly increasing in both arguments. Finally, Lemma 9 and the validity of  $M$  imply  $\zeta_{\gamma'}$  is strictly increasing.  $\square$

## References

- Aumann, R. (1976): “Agreeing to Disagree,” *Annals of Statistics*, 4, 1236–39.
- Bacharach, M. (1985): “Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge,” *Journal of Economic Theory*, 37, 167–90.
- Battigalli, P. and M. Siniscalchi (2003): “Rationalization and Incomplete Information,” *Advances in Theoretical Economics*, Vol. 3 No. 1, Article 3.
- Bergemann, D. and S. Morris (2007): “Strategic Distinguishability with an Application to Robust Virtual Implementation,” Cowles Foundation Discussion Paper 1609 Yale University.
- Blount, S. (1995), “When Social Outcomes Aren’t Fair: The Effect of Causal Attributions on Preferences,” *Organizational Behavior and Human Decision Processes*, 63, 131–44.
- Bolton, G. and A. Ockenfels (2000): “EEC - A Theory of Equity, Reciprocity and Competition,” *American Economic Review*, 90, 166–93.
- Brandenburger A. and E. Dekel (1993): “Hierarchies of Beliefs and Common Knowledge,” *Journal of Economic Theory*, 59, 1993, 189–98.
- Camerer, C. and A. Thaler (1995): “Ultimatums, Dictators and Manners,” *Journal of Economic Perspectives*, 9, 209–19.
- Cave, J. A. K. (1983): “Learning to Agree,” 12, 147–52.
- Charness, G. and M. Rabin (2002): “Understanding Social Preferences with Simple Tests,” *Quarterly Journal of Economics*, 117, 817–69.
- Cox, J. C., D. Friedman and S. Gjerstad (2004): “A Tractable Model of Reciprocity and Fairness”, mimeo, University of California, Santa Cruz.
- Dufwenberg, M. and G. Kirchsteiger (1999): “A Theory of Sequential Reciprocity,” *Games and Economic Behavior*, 47, 268–298.
- Dekel, E., D. Fudenberg and S. Morris (2006) “Topologies on Types,” *Theoretical Economics*, 1, 275–309.
- Dekel, E., D. Fudenberg and S. Morris (2006) “Interim Rationalizability,” *Theoretical Economics*, 2, 15–40.
- Ely, J. and M. Peski (2006) “Hierarchies of Beliefs and Interim Rationalizability,” *Theoretical Economics*, 1, 19–65.
- Falk A. and U. Fishbacher (1999): “A Theory of Reciprocity”, Working paper No. 6, University of Zurich.

- Falk A., E. Fehr and U. Fischbacher (2000): “Testing Theories of Fairness - Intentions Matter”, working paper no. 63. University of Zurich.
- Fehr E. and K. Schmidt (1999): “A Theory of Fairness, Competition and Cooperation.” *Quarterly Journal of Economics*, 114, 817–68.
- Geanakoplos J., D. Pearce and E. Stacchetti (1980): “Psychological Games and Sequential Rationality” *Games and Economic Behavior*, 1, pp. 60–80.
- Geanakoplos J., and H. Polemarchakis (1982): “We Can’t Disagree Forever,” *Journal of Economic Theory*, 28, 192–200.
- Levine, D, (1998): “Modeling Altruism and Spitefulness in Game Experiments,” *Review of Economic Dynamics*, 7, 348–52.
- Mertens J.F. and S. Zamir (1985): “Formulation of Bayesian Analysis for Games with Incomplete Information,” *International Journal of Game Theory*, 14, 1–29.
- Parikh, R. and P. Krasucki, (1990): “Communication, Consensus, and Knowledge,” *Journal of Economic Theory*, 52, 178–89
- Rabin, M., (1993): “Incorporating Fairness into Game Theory and Economics,” *American Economic Review*, 83, 1281–302.
- Segal, U. and J. Sobel (2004): “Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings,” Discussion Paper, University of California, San Diego.
- Sobel, J., (2004) “Interdependent Preferences and Reciprocity”, mimeo, University of California, San Diego.