Interdependent Preference Models as a Theory of Intentions[†]

Faruk Gul
and
Wolfgang Pesendorfer

Princeton University

July 2010

Abstract

We provide a preference framework for situations in which "intentions matter." A behavioral type describes the individual's observable characteristics and the individual's personality. We define a canonical behavioral type space and provide a condition that identifies collections of behavioral types that are equivalent to components of the canonical type space. We also develop a reciprocity model within our framework and show how it enables us to distinguish between strategic (or instrumental) generosity and true generosity.

[†] This research was supported by grants SES9911177, SES0236882, SES9905178 and SES0214050 from the National Science Foundation. We thank Stephen Morris and Phil Reny for their comments.

1. Introduction

In many economic settings, knowing the physical consequences of the interaction is not enough to determine its utility consequences. For example, Blount (1995) observes that experimental subjects may reject an unfair division when another subject willingly proposes it and yet might accept it when the other subject is forced to propose it. Hence, individuals care not just about physical consequences but also about the intentions of those around them. In this paper, we develop a framework for modeling intentions and how they affect others' behavior.

We call our descriptions of intentions *interdependent preference models* (IPMs). In an IPM, a person's ranking of social outcomes depends on the characteristics and personalities of those around him. Characteristics are attributes such as the individual's wealth, education or gender. Personalities describe how preferences respond to the characteristics and personalities of others. Thus, the personality defines a person's altruism, his desire to conform, his willingness to reciprocate or his inclination to be spiteful. To understand how our theory works, consider the following example.

Two individuals are to share a fixed sum of money. There are three possible outcomes: the sum of money can be given to one or the other person or it can be shared equally. There are two possible preferences for each player; either the player is selfish (S) and ranks getting the whole sum above sharing it equally or the player is generous (G) and ranks sharing the sum above getting it all. Giving the whole sum to the opponent is always the least preferred outcome.

There are 3 possible types for each person. Each type has the same characteristic and therefore types differ in their personalities only. The nicest type, 3, is generous irrespective of the opponent's type. Type 1 is the least nice type and is generous only if the opponent is type 3. Finally, type 2 is generous to all types other than type 1. The table below summarizes the mapping from type profiles to preference profiles:

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

Generous or Selfish

We call such a table an IPM. Levine (1998) introduces the first example of an IPM and uses it to address experimental evidence in centipede, ultimatum and public goods experiments.

Three features of IPMs are noteworthy: first, a type describes relevant personality attributes rather than information. These attributes determine both the person's and his opponent's preferences over outcomes, not their beliefs over an uncertain state of nature. To put it another way, IPM's do not incorporate asymmetric information (or interactive knowledge); they only model interactive preferences. Each entry in the table describes the preference the two individuals would have if they knew the other's type; the IPM does not address the question whether an individuals knows the others' type.

Second, an IPM does not describe the available strategic choices; it is not a game. We can study how Persons I and II above would play many different games. We can also use this IPM as the preference model for a competitive economy. Hence, IPMs describe only the preference environment not the institutional setting.

Third, in an IPM, individuals have preferences over *physical outcomes* and these preferences depend on the persistent personalities and characteristics of everyone involved, not on observed or predicted behavior or beliefs. Hence, the interaction of these fixed personalities determines whether each person is generous or selfish. Whether or not a person *acts* selfishly on a given day or believes the other will act selfishly is relevant only to the extent that these actions affect the physical outcome.

The last two observations highlight the main differences between IPMs and existing models of reciprocity. In our approach, there is a clear separation between the underlying preference framework (i.e., the IPM) and the particular institution (game, market). Geanakoplos, Pearce and Stacchetti's (1989) and the many reciprocity models based on their approach such as Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Segal and Sobel (2007) and Battigalli and Dufwenberg (2009), model preference interactions through a circular definition that permits preferences to depend on the very behavior (or beliefs about behavior) that they induce.

This circularity enables psychological games to accommodate many departures from standard theory. Some of these departures fall outside the reach of IPMs. For example, Geanakoplos, Pearce and Stacchetti show how psychological games can model a preference for being surprised or a preference for pandering to others' expectations. Such preferences are best described within the context of the particular interaction and hence are difficult to model as IPMs.

Because IPMs afford a separation between preferences and institutions they are well suited for analyzing economic design problems or for comparisons of institutions. Consider, for example, the (complete information) implementation problem: let f be a social choice rule (or performance criterion) that associates a set of outcomes with each profile, θ, of individual attributes. When is it possible to find a game form g such that all equilibrium outcomes of the game (g, θ) are among the desirable outcomes $f(\theta)$ for every θ? Note that the separation of the preference framework (i.e., the attributes θ) from the game form is essential for analyzing such a problem. Without this separation, just stating the implementation problem becomes a formidable task.

More generally, separating preferences from institutions is essential anytime we wish to evaluate a particular institution or assess a specific environmental factor: does the English auction Pareto dominate the Dutch auction? Do prohibitions on resale enhance efficiency? Does greater monitoring increase effort? Questions such as these demand a performance criterion over consequences that can be expressed without reference to specific institutions.

In an IPM, a type maps the other person's types to preference profiles. Hence, the definition of a type is circular. Our first objective is to identify a criterion to determine when this circularity is problematic and when it is not; that is, to identify when interdependent preference types can be reduced to preference statements. To see how this can be done, note that in the above example, the preference profile associated with type 3 does not depend on the opponent's personality and therefore type 3 can describe himself without any reference to his opponent's type. But once type 3 is identified, type 2 can describe himself as a personality that is generous only to the personality that has just been identified. Finally, type 1 can identify himself as someone who is generous to the two personalities that have been identified so far and no one else. Hence, in the above example, we can eliminate the circularity by restating each type as a hierarchy of preference statements.

We define the canonical type space for interdependent preferences as sequences of such hierarchical preferences statements. In Theorem 1, we show that each component of the canonical type space is an IPM. However, not every IPM is part of the canonical type space. Theorem 2 provides a simple condition on IPMs, (*validity*) that guarantees that an IPM is a component of the canonical type space. When a model fails validity, types cannot be reduced to 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 if the model is valid (Theorem 3).

As an application of our model, we define reciprocity and identify a class of valid interdependent preference models with reciprocating types. To do this, we consider settings in which preferences can be ranked according to their kindness. That is, we identify each preference with a real number and interpret higher numbers as kinder preferences. For example, assume x_i is i's consumption and

$$V^r(x_1, x_2) = u(x_1) + ru(x_2)$$

is the utility that person 1 enjoys if r is his preference. The IPM specifies how types determine r. Let $\delta(t, t') \in [\underline{r}, \bar{r}]$ be the preference of type t when the opponent is type t'. A higher r is a *kinder* preference. Type t is *nicer than* type t' if t is kinder than t' to every opponent type; a type *reciprocates* if it is kinder to a nicer opponent. Hence, if δ is increasing in the first argument, higher types are kinder. Then, if δ is also increasing in the second argument, all types reciprocate. In Theorems 4 and 5 we characterize two simple classes of reciprocity models.

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, we discuss the relationship between our results and technically related issues in that literature. In particular, we discuss Bergemann and Morris (2009) 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

We assume that there is one other person whose type affects the decision-maker. The two-person setting simplifies the notation and the extension to the n-person setting is straightforward. A set of social outcomes A, a set of characteristics Ω and a collection of preferences $\mathcal{R}$ on A characterizes the environment. For example, a social outcome could be the quantity of a public good together with a division of its cost or a pair of individual consumption levels. A characteristic might specify a player's occupation or education. The triple $(A, \mathcal{R}, \Omega)$ describes the underlying economic primitives.

We assume that A and Ω are compact metric spaces and that $\mathcal{R}$ is a nonempty and compact set of continuous preference relations[1] on A. An interdependent preference model (IPM) is a triple $M = (T, \gamma, \omega)$ where T is the type space, γ is a function that assigns a preference to each pair of types, and ω is a function that identifies the characteristic of each type.

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

Example 1: Consider the following straightforward generalization of the example in the introduction: there are three possible outcomes $A = \{(1,0), (0,1), (1/2, 1/2)\}$ and two possible preferences for each player, $\mathcal{R} = \{S, G\}$, as described in the introduction. There are k types, $T = \{1, \ldots, k\}$, who share the same characteristic and

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

In the introduction, we discuss the case of $k = 3$.

Note that players face a symmetric set of social outcomes. We can easily extend the analysis to asymmetric situations at the cost of complicating the notation. An alternative (and simpler) way to incorporate asymmetry is to define $\gamma(t, t')$ as a *pair of preferences*, one for when the decision maker is assigned the role of agent 1 and another for when

[1] A continuous preference relation on A is a complete and transitive binary relation R such that the sets $\{y \in A \mid yRx\}, \{y \in A \mid xRy\}$ are closed subsets of A.

he is assigned the role of agent 2. With this modification, our model can be applied to asymmetric situations.

For the IPM (T, γ, ω) 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 (T, Γ, ω). In that case, it is understood that Γ satisfies (1) for some $\gamma : T \times T \to \mathcal{R}$.

The function $\gamma(t, \cdot) : T \to \mathcal{R}$ describes how the agent's preference changes as a function of the opponent's type. Hence, $\gamma(t, \cdot)$ represents an agent's *personality.* Note, however, that the type space T is not a primitive of the economic environment. Therefore, $\gamma(t, \cdot)$ cannot serve as a satisfactory definition of an agents' personality. To be meaningful, a personality must be expressed in terms of the primitives $(A, \Omega, \mathcal{R})$. Next, we describe how this can be done.

Types are hierarchies of preference statements. In round 0, each type reports a characteristic (the characteristic of the type). In rounds $n \geq 1$, each type reports a set of preference profiles. The preference profile (R, R') is part of the round n report if, given the opponent's report in all previous rounds, it is possible that the player has preference R and his opponent has preference R'. Theorem 1 shows that when a collection of such hierarchies satisfies a straightforward consistency condition, it is an IPM.

Before providing the formal definition, it is useful to illustrate the correspondence between behavioral types and hierarchies of preference statements in Example 1. All types in this example have the same characteristic and therefore there is nothing to report in round 0. In round 1, each type reports the set of possible preference profiles given his type. For types $1, \ldots, k-1$, the preference profile is either (G, G) or (S, S). Hence, types $1, \ldots, k-1$ can report that both players will have identical preferences and that both the generous and the selfish preference profile are possible. For type k (the most generous type) we have $k + t > k$ for all t. Therefore, the preference profile is (G, G) when any type is matched with k. Round 1 thus identifies the most generous personality (type k).

In round 2, each type reports two sets of preferences, one in response to the round-1 (opponent's) report $\{(G, G)\}$ and one in response to the round-1 report $\{(G, G), (S, S)\}$. In response to $\{(G,G)\}$ all types must report $\{(G,G)\}$. This follows from a basic consistency requirement: if a player reports a single possible preference profile (R,R') in round n, then, in all successive rounds, his opponent must report (R',R) (i.e., the same preference profile with the roles permuted). There are three possible responses to $\{(G,G),(S,S)\}$: type k reports $\{(G,G)\}$, types $\{2,\ldots,k-1\}$ report $\{(G,G),(S,S)\}$ and type 1 reports $\{(S,S)\}$. Therefore, round 2 identifies type 1 as the least generous personality.

Continuing in this fashion, round 3 identifies the second most generous personality (type $k-1$) and round 4 identifies the second least generous personality (type 2). After k rounds, such hierarchical preference statements reveal all personalities in Example 1. Note that agents report preference *profiles* rather than individual preferences. Individual's types place restrictions not only on their own preference but also on the preferences of their opponents. Hence, to permit the full generality of possible preference interactions, it is necessary that hierarchical statements convey (sets of) preference profiles not just their own preferences.[2]

To define personalities, we use the following notation and definitions. 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 $f: X\to\mathcal{H}_Z$ such that their graph $G(f)=\{(x,z)\in X\times Z \,|\, z\in f(x)\}$ is closed in $X\times Z$.[3] 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 $f: X\to Z$ with the function $\bar{f}: X\to\mathcal{H}_Z$ 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.

We let $\mathcal{H}$ denote $\mathcal{H}_{\mathcal{R}\times\mathcal{R}}$, the collection of all closed subsets of preference profiles (closed subsets of $\mathcal{R}\times\mathcal{R}$). Types are *consistent preference response hierarchies.* As illustrated above, these hierarchies specify *sets of preferences* that are gradually refined as more

[2] In example 1 both agents have the same preference hence statements about agents' own preferences are sufficient. However, this is not true in general. To identify certain personality types, it may be necessary to convey information about the opponent's preference.

[3] Hence, $\mathcal{C}(X,\mathcal{H}_Z)$ is the set of upper hemi-continuous correspondences from X to Z.

detail about the opponent's personality is revealed. We first define the hierarchies and then provide the appropriate consistency condition.

Definition: *A collection of nonempty compact sets* $(\Theta_0, \Theta_1, \ldots)$ *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 $\theta_1 = (\theta_0, f_1)$ specifies a characteristic θ_0 and a map $f_1 : \Theta_0 \to \mathcal{H}$ that associates each opponent characteristic with a set of preference profiles (the round 1 statements the example above). More generally, the entry θ_k consists of the previous entry (θ_{k-1}) and the function $f_k : \Theta_{k-1} \to \mathcal{H}$ that specifies for each θ_{k-1} of the opponent a set of possible preference profiles.

Not all preference response hierarchies identify meaningful personality types. To illustrate this, assume there is a single characteristic. Then, we can omit round 0 and f_1 is the set of possible preference profiles for a particular type. Assume that $f_1 = \{(R, R')\}$ and $f_2(f_1') \neq \{(R, R')\}$ for some $f_1' \in \Theta_1$. In that case, $\theta_2 = (f_1, f_2)$ is inconsistent: the round 1 report says that only (R, R') is possible while the round 2 report says that there is some opponent (f_1') that leads to a different profile. Conversely, suppose that $f_1 = \{(R, R'), (\hat{R}, \hat{R}')\}$ and $f_2(f_1') = \{(R, R')\}$ for all $f_1' \in \Theta_1$. In that case, the round 1 report says that two preference profiles are possible ((R, R') and $(\hat{R}, \hat{R}')$) but according to f_2, there is no f_1' for the opponent such that $(\hat{R}, \hat{R}')$ is a possible preference profile. Again, the type $\theta_2 = (f_1, f_2)$ is inconsistent. Therefore, the consistency of $\theta_2 = (f_1, f_2)$ requires

$$\bigcup_{f_1' \in \Theta_1} f_2(f_1') = f_1$$

Next, assume that $f_1 = \{(R, R')\} \in \Theta_1$ and $f_1' = \{(\hat{R}, \hat{R}')\} \in \Theta_1$. Then, both types report a single possible preference profile. Since these two types may be matched, it follows that those two preference profiles must coincide. Therefore, consistency requires that $(R, R') = (\hat{R}', \hat{R})$. (Recall that our notation omits player names; the first entry refers to the player and the second to the opponent.) More generally, the intersection of any pair f_1 and f_1' (with the entries permuted) must be non-empty.

The next definition below specifies the same two consistency requirements for all levels of the hierarchy.

Definition: *The system of preference response hierarchies* $(\Theta_0, \Theta_1, \ldots)$ *is consistent if:*

(i) For all $n \geq 1$, *for all* $(\theta_{n-1}, f_n, f_{n+1}) \in \Theta_{n+1}$, *and for all* $\bar{\theta}_{n-1} \in \Theta_{n-1}$

$$f_n(\bar{\theta}_{n-1}) = \bigcup_{\{f'_n \,|\, (\bar{\theta}_{n-1}, f'_n) \in \Theta_n\}} f_{n+1}(\bar{\theta}_{n-1}, f'_n)$$

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

$$(R, R') \in f_n(\theta'_{n-1}) \text{ and } (R', R) \in f'_n(\theta_{n-1})$$

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

Definition: *Let* $(\Theta_0, \Theta_1, \ldots)$ *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* Θ *is a component of behavioral types if* Θ *is compact and if for all* $\theta, \theta' \in \Theta$ *with* $\theta = (f_0, f_1, \ldots)$

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

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 partition of $\mathcal{F}$.

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

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

The function $\Psi(\theta, \cdot)$ is the *personality* of type θ. It describes how the player responds to different opponent personalities. Requirement (ii) in the definition of consistency ensures that the function ψ satisfies the following symmetry condition.

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

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

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

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

It follows from Theorem 1 that any component $\Theta \in \mathcal{I}$ is an (IPM): for a symmetric Ψ, there is a $\psi : \Theta \times \Theta \to \mathcal{R}$ such that $\Psi(\theta, \theta') = (\psi(\theta, \theta'), \psi(\theta', \theta))$. Then, since Θ is compact (by definition) and ψ is continuous (by Theorem 1), it follows that every component of the canonical type space is an IPM. We record this observation as a corollary.

Corollary: *If $\Theta \in I$, then (Θ, ψ, ω) is an IPM.*

3. Valid Models

Suppose the environment has a single characteristic and two possible preferences, a and b. Consider the following IPM:

	1	2
1	(a, a)	(b, b)
2	(b, b)	(a, a)

Table 3

This IPM is not an element of the canonical type space. To see why not, note that all types have the same characteristic, there is nothing to report in round 0. The round 1 set of possible preference profiles is $\{(a, a), (b, b)\}$ for both players. Since round 1 statements are identical for both types, all higher round statements must be identical as well. Hence, for all n the function f_n is constant (with value $\{(a, a), (b, b)\}$). The two types in this IPM have identical preference response hierarchies.

The IPM in table 3 illustrates a case in which the IPM introduces a distinction between types 1 and 2 that has no counterpart in terms of the model's primitives. Every preference statement that holds for type 1 is also true for type 2. Therefore, we cannot express the personalities of those two types as preference statements.

We can interpret the IPM in table 3 as an *incomplete model.* For example, it might be that the model omits a type.

	1	2	3
1	(a, a)	(b, b)	(b, a)
2	(b, b)	(a, a)	(a, a)
3	(a, b)	(a, a)	(a, a)

Table 4

In the IPM in table 4, type 3 always prefers a irrespective of the opponent's type. Type 2 accommodates this preference while type 2 does not. Hence, types 1 and 2 have different personalities, and can be distinguished through their response to type 3. The IPM depicted in table 3 could be interpreted as table 4 with an omitted type.

Alternatively, it might be that the model has omitted a characteristic: if types 1 and 2 have different characteristics (type 1 is wealthy, type 2 is poor) then the IPM in table 3 is a component of the canonical type space: both types have preference a if the opponent's has the same wealth and preference b otherwise.

The two interpretations have obviously very different implications for applications. If players' behavior depends on the opponent's wealth (an omitted characteristic) then the observability of wealth is a key determinant of outcomes. If players' behavior depends on their response to some third personality type (type 3) then observability of wealth should have no effect and instead the observability of past play will affect outcomes. Since the IPM is not a component of the canonical type space, we cannot express personalities in terms of the underlying primitives of the model. As a result, we cannot determine how information about the underlying primitives affects behavior.

Next, we provide a criterion (validity) that identifies whether or not an IPM is a component of the canonical type space. A partition $\mathcal{D}$ of T is a pairwise disjoint collection of non-empty subsets such that $\bigcup_{D \in \mathcal{D}} D = T$. Let D^t denote the unique element of $\mathcal{D}$ that contains t. The partition $\mathcal{D} = \{\{t\} \mid t \in T\}$ is called the *finest* partition. Let (T, γ, ω) be an IPM. Recall that

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

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

Definition: *The IPM (T, γ, ω) is valid if the finest partition of T is the only partition $\mathcal{D}$ that satisfies*

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

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

Validity requires that it be impossible to partition the type space in a manner that yields a partition element with multiple (indistinguishable) types. Theorem 2 shows that any valid IPM corresponds to a component $\Theta \in \mathcal{I}$. Two IPM's (T, γ, ω), (T', γ', ω') are isomorphic if there exists a homeomorphism $\iota : T \to T'$ 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 (T, γ, ω) is valid if and only if it is isomorphic to a component of the canonical type space.*

We have interpreted the invalid IPM above as an incomplete model, either missing types or missing characteristics. Note that any finite invalid IPM can be "validated" by adding new types or new characteristics. Any model is obviously valid if each type has a distinct characteristic. For the missing types interpretation, it is easy to show that any finite, invalid IPM can be embedded in a valid IPM with a larger type space. More precisely, assume that $\mathcal{R}$ contains at least 2 preferences and consider a finite IPM (T, γ, ω). Then, there exists a valid IPM $(\hat{T}, \hat{\gamma}, \hat{\omega})$ such that $T \subset \hat{T}, \gamma(t) = \hat{\gamma}(t)$ for all $t \in T$ and $\omega(t) = \hat{\omega}(t)$ for all $t \in T$.

4. Validity and Communicability

We have interpreted preference response hierarchies as players' conditional preference statements that gradually reveal their type. Those hierarchies impose a particular protocol of how these statements unfold. In this section, we show that no other protocol can do better. More formally, we introduce a general framework for communication and show that agents can reveal their types through communication if and only if the IPM is valid. Hence, types are communicable if and only if the IPM is a component of the canonical type space.

For simplicity, we assume that there is single characteristic[4] and a finite number of behavioral types. To formulate a model of communication, we need to translate the IPM into an epistemic model. An epistemic model is a finite set of states S, a map $\nu : S \to \mathcal{R} \times \mathcal{R}$ that associates a preference profile with each state and a pair of partitions $\mathcal{T}_1$ and $\mathcal{T}_2$ of S that represent the players' knowledge. Elements of $\mathcal{T}_i$ are player i's epistemic types.

Let (T, γ) be an IPM with a single characteristic and a finite set of types. In the equivalent epistemic model, each player knows his own type and knows nothing about his opponent's type; that is, any type in T possible. Formally, the epistemic model $E = \{S, \mathcal{T}_1, \mathcal{T}_2, \nu\}$ is *equivalent to* the IPM (T, γ) if there is a bijection $\zeta_i : T \to \mathcal{T}_i$ such that $\{\Gamma(t, t')\} = \nu(\zeta_1(t) \cap \zeta_2(t'))$ for all $t, t' \in T$. The bijection ζ_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).*

[4] Extending the analysis below to IPMs with multiple characteristics is straightforward. We assume a single characteristic to keep the notation simple.

A collection of subsets $\mathcal{K}$ of a set S is an algebra (or equivalently, a *language*) if it contains S 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 $\mathcal{K} \wedge \mathcal{L}$ be the largest that is contained in both.

We call each $A \in \mathcal{L}$ is a *word.* The primitives of an interactive preference model are preference statements. Hence, the relevant language is the *the language of preferences.*[5] A subset A of S is a word in the language of preferences if $A = \{s \in S \,|\, \nu(s) \in V\}$ for some set of preference profiles $V \in \mathcal{H}$.

Definition: *The language of preferences is* $\mathcal{L}_p = \{A \subset S \,|\, A = \nu^{-1}(V), V \in \mathcal{H}\}$.

To illustrate these definitions, we apply them to Example 1 from the previous section.

Example 1′: (epistemic form) Let $T = \{1, \ldots, k\}$ and define γ as in Example 1 above:

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

Then, define $\{S, \mathcal{T}_1, \mathcal{T}_2, \nu\}$ equivalent to (T, γ) as follows: let $S = \{(i,j) \,|\, i \in T, j \in T\}$ and $\nu(i,j) = (\gamma(i,j), \gamma(j,i))$ for all $(i,j) \in S$. The information partitions are $\mathcal{T}_1 = \{\{(i,1), \ldots, (i,k)\} \,|\, i \in T\}$ and $\mathcal{T}_2 = \{\{(1,i), \ldots, (k,i)\} \,|\, i \in T\}$.

There are two possible preference profiles, $\{G, G\}$ and $\{S, S\}$. Therefore, the language of preferences has three non-empty words: $A = \{(i,j) \,|\, i+j > k\}$, $B = \{(i,j) \,|\, i+j \leq k\}$ and $S = A \cup B$. The word A corresponds to $\{(G,G)\}$, B corresponds to $\{(S,S)\}$ and $S = A \cup B$ corresponds to $\{(S,S), (G,G)\}$.

A person with knowledge $\mathcal{T}_i$ can use 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 $\{\mathcal{T}_i * A, \mathcal{T}_i * (S \backslash A)\}$ to j. These are words derived from A and $S \backslash A$ that i *understands*; that is, at every $s \in S$, i knows whether

[5] The formalism developed here can be used to define communicability with respect to any given language.

or not $\mathcal{T}_i * A$ and $\mathcal{T}_i * (S\backslash A))$ applies (i.e., is true); he knows whether or not he knows A and he knows whether or not he knows $S\backslash A$. Then, using standard logical operations he can also communicate other words such as $[S\backslash(\mathcal{T}_i * A)] \cap [S\backslash(\mathcal{T}_i * (S\backslash A))]$; i.e., that he knows neither A nor $S\backslash A$. We let

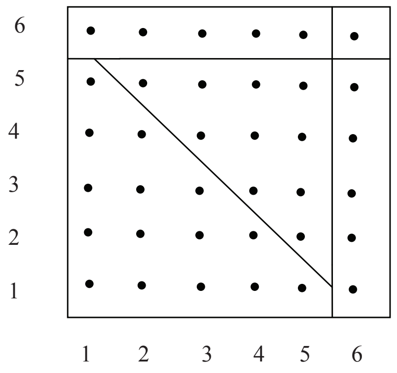
$$\mathcal{T}_i * \mathcal{L}$$

denote the language that can be communicated in this way. That is, $\mathcal{T}_i * \mathcal{L}$ is the smallest algebra that contains $\mathcal{T}_i * A$ for every $A \in \mathcal{L}$.[6]

Let Λ be the collection of all algebras and let $\Lambda_p := \{\mathcal{L} \in \Lambda \,|\, \mathcal{L}_p \subset \mathcal{L}\}$.[7] Define the function $F : \Lambda_p \to \Lambda_p$ as follows:

$$F(\mathcal{L}) = \mathcal{L} \vee [\mathcal{T}_1 * \mathcal{L}] \vee [\mathcal{T}_2 * \mathcal{L}]$$

Hence, $F(\mathcal{L})$ is the refinement of $\mathcal{L}$ that results from one round of communication. We can describe the function F for the example above. Let $\mathcal{L}$ be the language of preferences consisting of the words $A = \{(i,j) : i + j > k\}, B = \{(i,j) : i + j \leq k\}$ and $A \cup B$. Then, the partition in the figure below generates the algebra $F(\mathcal{L})$. As the figure shows, the word B remains unchanged whereas the word A is partitioned into 3 separate words.



[6] 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.

[7] The set Λ is a lattice under the binary relation $\subset$ and Λ_p is a sublattice.

Let $\mathcal{L}^1 = \mathcal{L}_p$ and inductively define $\mathcal{L}^{n+1} = F(\mathcal{L}^n)$. Players can continue communicating until they reach a fixed point of F. Since S is finite, there is n such that $F(\mathcal{L}^n) = \mathcal{L}^n$. Let $\mathcal{L}^*$ denote this fixed point. Through this process, players can communicate the algebra

$$\mathcal{C}_p := [\mathcal{T}_1 * \mathcal{L}^*] \vee [\mathcal{T}_2 * \mathcal{L}^*]$$

Then, a collection of subsets $\mathcal{M}$ of S is communicable if only if $\mathcal{M} \subset \mathcal{C}_p$.

In the sequence $\mathcal{L}^1, \mathcal{L}^2, \ldots$ players communicate all information in each round. Consider any other sequence $\hat{\mathcal{L}}^1, \hat{\mathcal{L}}^2, \ldots$ starting at the same initial point $\hat{\mathcal{L}}^1 = \mathcal{L}_p$ such that in 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. This process will refine the language until a fixed point $\hat{\mathcal{L}}$ of F is reached. We claim that $\hat{\mathcal{L}} = \mathcal{L}^*$. To see this, observe that F is a monotone function, that is,

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

Applied to the fixed point $\hat{\mathcal{L}}$, the equation above implies $F(\mathcal{L}) \subset \hat{\mathcal{L}}$ if $\mathcal{L} \subset \hat{\mathcal{L}}$. Since $\mathcal{L}^1 \subset \hat{\mathcal{L}}$, we conclude that $\mathcal{L}^n \subset \hat{\mathcal{L}}$ for all n and therefore $\mathcal{L}^* \subset \hat{\mathcal{L}}$. That $\hat{\mathcal{L}} \subset \mathcal{L}^*$ follows from the fact that $\hat{\mathcal{L}}^n \subset \mathcal{L}^n$ for all n (by construction). Hence, $\mathcal{L}^* = \hat{\mathcal{L}}$ and $\mathcal{L}^*$ is the outcome any communication protocol. A particular collection of subsets is communicable in $\mathcal{L}_p$ if this collection is contained in $\mathcal{C}_p$. We wish to characterize when players' types are communicable.

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

An IPM is communicable if players can express their types in the language of preferences. Theorem 3 below shows that validity identifies exactly those models that can be communicated.

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.*

We interpret communicability as a test that a well-defined, self-contained model of interdependent preferences types must satisfy: responding to the opponent's personality requires understanding his personality and each player's understanding is the sum of what he knows at the outset (i.e., his type) and what he can learn through communication with the other player.[8] A model that violates this communicability test can, at best, be interpreted as an incomplete model: players respond to personality types that are well defined in a larger context (in a richer model) but ill-defined in the IPM at hand.

5. Reciprocity

A reciprocating personality[9] is one that is kinder to nicer opponents. Hence, in our formal definition of reciprocity, we assume an exogenous "kindness" ranking on preferences. The literature often assumes that individuals have exogenously specified selfish utilities and identifies altruism (or generosity or kindness) with the relative weight a person puts on other's selfish utility. For example, let $(x, y) \in A = [0, 100] \times [0, 100]$ be the vector consumptions and assume that selfish utilities are linear. Then, let U^r such that

$$U^r(x, y) = x + ry$$

be the utility function representing preference with parameter $r \in [-1, 1]$. A natural kindness order on these preferences is the "$\geq$" ranking of their parameters. That is, the preference (with parameter) r is kinder than the preference r' if and only if $r \geq r'$.

More generally, we assume that there is a continuous one-to-one function $\tau : \mathcal{R} \to I\!R$ and interpret $\tau(R) \geq \tau(R')$ to mean R is kinder than R'. We say that M is an ordered IPM when such τ exists. When the IPM is ordered, it is convenient to identify each preference R with $\tau(R)$ and suppress preferences. Then, we let $\delta = \tau \circ \gamma$ and refer to (T, δ) as an ordered IPM.

Definition: *In an ordered IPM, type t is nicer than type t' if $\delta(t, t'') \geq \delta(t', t'')$ for all t''; type t reciprocates if $\delta(t, t') \geq \delta(t, t'')$ whenever t' is nicer than t''. An ordered IPM is a reciprocity model if it is valid, the niceness relation is complete and every type reciprocates.*

[8] Since we wish to identify what can be understood in principle, we are ignoring incentives. In practice, an agent may learn less than $\mathcal{C}_p$ since his opponent may strategically withhold information.

[9] For simplicity, we assume throughout this section that all types have the same characteristic and hence use type and personality interchangeably.

This definition incorporates both positive and negative reciprocity.[10] Let r^* be a reference level of fairness and let

$$\delta(e(t), t) = r^*$$

Type t's opponent exceeds the reference level of fairness whenever $t' \geq e(t)$. So, type t exhibits *positive reciprocity* if $\delta(t, \cdot)$ is flat or nearly flat when $t' < e(t)$ but steep when $t' > e(t)$. Thus, type t reciprocates when opponent types are nicer than the threshold $e(t)$ but does not reciprocate when opponent types are below $e(t)$. Conversely, type t exhibits *negative reciprocity* if $\delta(t, \cdot)$ is steep when $t' < e(t)$ but flat when $t' > e(t)$. Thus, type t reciprocates when opponent types are less nice than the threshold $e(t)$ but does not reciprocate when opponent types are above $e(t)$.

The simplest kind of an ordered IPM is one in which types are also real numbers. In such a model, if δ is increasing in the first argument, then bigger types are nicer. If it is also increasing in the second argument, then all types reciprocate. Theorem 4 shows that adding a mild genericity condition to such a model ensures validity. The theorem also shows that, in fact, all reciprocity models are of this kind.

Definition: *The ordered IPM (T, δ) is simple if $T \subset I\!R$; a simple IPM is increasing if δ is nondecreasing in both arguments and $\delta(t, \cdot) = \delta(t', \cdot)$ implies $t = t'$.*

Theorem 4: *An ordered IPM is a reciprocity model if and only if it is isomorphic to some simple increasing IPM.*

To prove Theorem 4, we first show that validity together with the continuity and compactness properties of IPMs ensure that the types in a reciprocity model can be identified with real numbers. For the converse, we show that if δ is nondecreasing in both arguments, then the simple IPM (T, δ) is valid if and only if δ satisfies the genericity condition $(t, \cdot) = \delta^*(t', \cdot)$ implies $t = t'$. Hence, the nondecreasingness of δ ensures that each type in a simple model is reciprocating and given the nondecreasingness, the genericity condition is exactly what is needed for a simple model to be valid.

Next, we examine the special case of an ordered IPM with two possible preference profiles. Hence, we identify $\mathcal{R}$ with $\{0, 1\}$ where 1 is the generous (i.e., kinder) preference,

[10] Dohmen, Falk, Huffman and Sunde (2009) show that positive reciprocating people (i.e., types) enjoy better job market outcomes and more life satisfaction than negative reciprocating types.

0 is the selfish preference and assume that $\delta(t, t') = (0, 0)$ or $\delta(t, t') = (1, 1)$ for all $t, t' \in T$. We call such models *binary* IPMs.

As we show in Theorem 5 below, there are only two classes of valid binary IPMs. Both classes consist of simple increasing IPMs and hence are reciprocity models. In the first class, the highest type is generous irrespective of the opponent's type and all opponents are generous when matched with the nicest type. Example 1 belongs to this class. Let

$$\delta_m(i, j) = \begin{cases} 1 & \text{if } i + j > m \\ 0 & \text{otherwise} \end{cases}$$

Let $K = \{1, \ldots, k\}$ be the set of types. Then, $M_k^0 = (K, \delta_k)$ is the first class of binary IPMs. In the second class, the lowest type is always selfish irrespective of the opponent's type and all opponents are selfish when they are matched with the lowest type. This class of binary IPMs is $M_k^1 = (K, \delta_{k+1})$ for $k = 1, \ldots$. Theorem 5 shows that these are the only valid binary IPMs.

Theorem 5: *A binary IPM is valid if and only if it is isomorphic to some M_k^i.*

It is easy to verify that every M_k^i is a reciprocity model. Hence, Theorem 5 establishes that valid IPMs are reciprocity models. To gain intuition for Theorem 5, first note that, by compactness and continuity, a binary reciprocity model must have a finite number of types $\{1, \ldots, m\}$. 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 preference profiles, validity ensures that $\delta(t, \cdot)$ is constant for some type t. Suppose this constant is 1 and without loss of generality let $t = k + 1$. Then, we show that the validity of $(\{1, \ldots, k + 1\}, \delta)$ implies that $(\{1, \ldots, k\}, \delta)$ is valid and that there exists no $t \leq k$ such that $\delta(t, \cdot) = 1$. Then, by the inductive hypothesis, the restriction of δ to $\{1, \ldots, k\}$ must be δ_{k+1}, implying that $\delta = \delta_{k+1}$.

5.1 Modeling Intentions with IPMs

As we noted in the introduction, IPMs are not adequate for modeling all departures from the standard framework. By the standard framework, we mean what Falk and Fischbacher (2006) call the "consequentialistic perspective," that is, any model in which the physical description of outcomes is sufficient for identifying utility outcomes. Even some forms of reciprocity may fall outside of the reach of IPMs. For example, a player may care only about whether or not another player *acted* generously and not about the other player's (persistent) personality.[11]

However, when modeling intentions, identifying persistent attributes as the carriers of utility has some advantages. We considered one of those advantages in the introduction: the implied separation of preferences from institutions facilitates the analysis of economic design problems. Here, we will discuss a second advantage: it enables reciprocity models to differentiate between *acting* generously out of self-interest and genuine kindness.

Berg, Dickhuat and McCabe (1995) provide experimental evidence indicating that (i) subjects often act generously towards others and trust them to reciprocate and (ii) this trust/generocity is often rewarded. Berg, Dickhuat and McCabe consider factors that might facilitate such trust and reciprocity. Our goal in this subsection is to suggest a novel experiment that would enable the experimenter to determine if this trust/generocity reflects genuine concern for the opponent or if it is strategic and motivated by the expectation of reciprocity. Then, we show how our theory might be useful for organizing the results of such an experiment.

Consider the following game: player 1 is either generous/trusts (g) player 2 or he does not (s). Afterwards, nature chooses either 0 or 1. If nature chooses 0, the game ends. The outcome $(g, 0)$ yields $(70, 30)$; that is, 70 dollars for player 1 and 30 dollars for player 2 while $(s, 0)$ yields $(80, 0)$. If nature chooses 1, then player 2 chooses an action; she either accepts player 1's decision (a) or declines it (d). The outcomes $(g, 0)$ and $(g, 1, a)$ yield the monetary payoffs $(70, 30)$ while $(s, 0)$ and $(s, 1, a)$ yield the monetary payoffs $(80, 0)$. If player 2 chooses d, she gets 40 dollars and player 1 gets 0 dollars. That is, $(g, 1, d)$ and $(s, 1, d)$ both yield the monetary payoffs $(0, 40)$. Let $\alpha \in (0, 1)$ be the probability that nature chooses 1.

Consider how changing α might affect behavior. First, let α be close to 1. Then, if player 1 chooses s and nature chooses 1, player 2 is likely to choose d: by choosing s, player 1 has shown no generocity/trust and therefore player 2 is likely to be ungenerous as well. Knowing this, player 1 believes he is likely to get 0 if he chooses s. Hence, when α is high,

[11] Falk and Fischbacher's (2006) definition of reciprocity reflects this view.

player 1 will be inclined to choose g even if he puts no weight 2's well-being. Thus, player 1's generocity/trust in this case may be strategic.

In contrast, when α is close to zero, the action g reveals a genuinely generous player 1; had he chosen s, he would have (almost) guaranteed himself 80 and he is giving this up for player 2's benefit. Therefore, conditional on the node $(g, 1)$ being reached, we would expect player 2 to be more inclined to honor player 1's trust when α is low than when it is high.

Thus, when α is close to 0, g is proof of player 1's good intentions (i.e., that he is type 2) and is fully rewarded. When α is close to 1, g can mean that player 1 has good intentions or that he would like player 2 to think he has good intentions. The possibility that player 1's generosity is not genuine should make player 2's less reciprocating.

To see how a reciprocity model can match the intuition outlined above, we will model preferences with the binary IPM M_2^0 defined in the previous section. To be concrete, let $r \in \{0, 1\}$ be the preference that U^r below represents:

$$U^r(x, y) = x + ry$$

Then, consider the IPM $(\{1, 2\}, \delta)$, where

$$\delta(t, t') = 1 \text{ if and only if } t + t' > 2.$$

Assume that both players are drawn from the same population with 10% type 2's. For α sufficiently small, the game above has a unique equilibrium: player 1 chooses g if and only if he is type 2 and player 2 always accepts g and accepts s if and only if she is type 2. For α close to 1, the game again has a unique equilibrium: player 1 chooses g for sure if he is type 2 and randomizes between g and s if he is type 1. Player 2 accepts g for sure if she is type 2 and randomizes between a and d if she is type 2.

These two equilibria match the intuition above exactly: if α is close to 0, choosing g is unambiguously generous and such generosity is rewarded. If α is close to 1, both generosity and self-interest are possible motives for choosing g and therefore player 2's response is more qualified; sometimes she reciprocates and sometimes she doesn't.

With interdependent preferences, different game forms create different incentives to reveal (or conceal) intentions. Understanding the behavioral consequences of a particular institution or environmental factor, then, amounts to understanding the incentives it creates for signalling intentions.

6. Related Literature on Belief (or Possibility) Hierarchies

That each types can be identified with a unique hierarchy of preference statements is a central property of our model. The same objective – relating differences in types to differences in payoff relevant primitives – can also be pursued when types are exogenous parameters.[12] Bergemann and Morris (2007), (2009) and Bergemann, Morris and Takahashi (2010) establish that this question plays a central role in implementation theory. Bergemann and Morris (2007) permits asymmetric information (i.e., players form conjectures over their opponents' preference types) and they prove two results in which conditions similar to validity play a role.

To facilitate the comparison, we consider finite, symmetric two-person IPMs with a single characteristic.[13] Each pair of types yields a pair of von Neumann-Morgenstern utilities on Z, the set of all lotteries over outcomes. Let (T, γ) 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 $g : A_1 \times A_2 \to Z$. Hence, A_i is player i's pure strategy set and g is the outcome function that relates pure strategy profiles to lotteries over outcomes. Bergemann and Morris make a mild genericity assumption: given any belief over opponent types and actions, no type is ever indifferent over all of his own actions.

Bergemann and Morris define rationalizable actions as follows: each round, players are allowed any conjecture over opponent types and allowed actions for those types. Then, actions that are never best responses for a type against all such conjectures are eliminated and become no longer unavailable for that type. Actions that are never eliminated are rationalizable. Bergemann and Morris (2007) call two types strategically equivalent if

[12] 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 possibility hierarchy. The latter concept is due to Mariotti, Meier and Piccione (2005).

[13] Bergemann and Morris (2007) allow for arbitrary finite $n-$person IPMs.

they have the same set of rationalizable actions in every game. To relate their Proposition 5 to our analysis of communicability, we present the following stronger notion of validity:

Definition: *The IPM (T, γ, ω) is strongly valid if the finest partition of T is the only partition $\mathcal{D}$ that satisfies*

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

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

The difference between validity and strong validity is that γ replaces Γ 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 (2007) as follows:

Proposition: *If (γ, T) satisfies the genericity condition above and fails strong validity, then there are 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 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, there should be some 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 utility only if both have the same kind of radio 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.[14]

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) whenever a mild genericity condition is satisfied. 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,[15] there are more than two communicating agents and explicit, general protocols determining who speaks when.[16]

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

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

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

[14] 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.

[15] See for example, Cave (1983) and Bacharach (1985).

[16] See Parikh and Krasucki (1990).

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:

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

Thus, an agent who knows the event E, knows that the true preference profile is in $g(E)$. This 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.

7. Appendix

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

$$\underline{\lim} A_n = \{z \in Z \,|\, z = \lim z_n \text{ for some sequence } z_n \text{ such that } z_n \in A_n \text{ for all } n\}$$
$$\overline{\lim} A_n = \{z \in Z \,|\, z = \lim z_{n_j} \text{ for some sequence } z_{n_j} \text{ such that } z_{n_j} \in A_{n_j} \text{ for all } j\}$$

Let X be a metric space and $p : X \to \mathcal{H}_Z$. 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 Pearcy (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. □

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. □

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 $\overline{\lim} G(p_n) \subset G(p) \subset \underline{\lim} G(p_n)$. □

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

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

Lemma 4: *Let $\hat{\theta} \in \Theta \in \mathcal{I}$ with $\hat{\theta} = (f_0, f_1, \ldots)$ 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

$\theta = (\theta_{n-1}, g_n, g_{n+1}, \ldots) \in \Theta$ 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 $q : X \times Y \to Z$ 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 $k \in I\!N$, $m \geq k$ 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,

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

$\square$

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 $p^{-1} : p(X) \to X$ 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. $\square$

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

7.1 Proof of Theorem 1:

We first show that ϕ is continuous. Consider any sequence $\theta^k = (f_0^k, f_1^k, \ldots) \in \Theta$ such that $\lim \theta^k = \theta = (f_0, f_1 \ldots) \in \Theta$. Let $\phi(\theta) = (f_0, f)$ and $\phi(\theta^k) = (f_0^k, f^k)$ for all k. Let $\theta^k = (f_0^k, f_1^k, \ldots)$, $\theta = (f_0, f_1, \ldots)$ 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 $f_N^k \to f_N$ 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 ϕ is continuous.

Next, we prove that ϕ is one-to-one. Pick any $(f_0, f_1, \ldots), (g_0, g_1, \ldots) \in \Theta$. Let $(f_0, f) = \phi(f_0, f_1, \ldots)$ and $(g_0, g) = \phi(g_0, \ldots)$. 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 ϕ is continuous and one-to-one and Θ is compact, it follows from Lemma 6 that ϕ is a homeomorphism from Θ to $\phi(\Theta)$. The continuity of Ψ follows from the compactness of Θ and Lemma 5. □

7.2 Proof of Theorem 2:

We say that $\mathcal{D}$ is strongly continuous if the function $\sigma : X \to \mathcal{D}$ 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 partitions $\mathcal{D}_n$ on T as follows:

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

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

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

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

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

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 $\mathcal{D}_n$ is 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 $\mathcal{D} \neq \{\{t\} \,|\, t \in T\}$, then M is not valid. Suppose M is not valid and hence there exists a continuous partition $\mathcal{D}^*$ that satisfies (ii) in the definition of validity, other than the finest partition. 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 $\mathcal{D}^*$ satisfies (ii) in the definition of validity 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

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

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 $f_n^t : \Theta_{n-1} \to \mathcal{H}, \Theta_n, \iota_n : T \to \Theta_n$ as follows:

$$f_n^t(\theta_{n-1}) = \Gamma(t, \iota_{n-1}^{-1}(\theta_{n-1}))$$

$$\iota_n(t) = (\iota_{n-1}(t), f_n^t)$$

$$\Theta_n = \iota_n(T)$$

Let

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

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

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

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

Fact 1: *For all n, the functions ι_n are onto and continuous and the sets Θ_n are non-empty and compact.*

Proof: We will prove inductively that Θ_n are nonempty, compact, ι_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 Θ_{n+1} is compact. The functions ι_n is onto by definition. □

Fact 2: *(i) The function ι is onto. (ii) $\iota_n(t) = \iota_n(s)$ if and only if $D^t_n = D^s_n$. (iii) $f^t_n(\iota_{n-1}(s)) = g^t_n(s)$.*

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

Next, we prove that $\iota_n(t) = \iota_n(s)$ if and only if $D^t_n = D^s_n$. To see this, note that for $n = 0$, the assertion is true by definition. Suppose, it is true for n. Then, if $s \in D^t_{n+1}$, we have $s \in D^t_n$ and $\Gamma(t, D_n) = \Gamma(s, D)$ for all $D \in \mathcal{D}_n$. Hence, $f^t_{n+1} = f^s_{n+1}$ 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 $f^t_{n+1} = f^s_{n+1}$ and $i^t_n = f^s_n$. Therefore,by the inductive hypothesis, $s \in D^t_{n+1} \in \mathcal{D}_{n+1}$.

Part (iii) follows from part (ii) and the definitions of g^t_n, f^t_n. □

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

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 $d(g^t, g^t_N) < \epsilon$ and $\bar{d}(g^t_N) < \epsilon$. By Lemma 1(ii) $g^{t_n}_N \to g^t_N$. By Lemma 7(i) we can choose m so that $\bar{d}(g^{t_k}_N) \leq 2\epsilon$ for all $n \geq m$. Therefore, by Lemma 7(ii), $d(g^{t_n}_n, g_n) < 3\epsilon$ for all $n > \max\{m, N\}$. It follows that $d(g^{t_n}_n, g) < 4\epsilon$ for all $n > \max\{m, N\}$ as desired. □

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

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

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

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

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

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

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

To prove that ι is a homeomorphism, we prove that ι 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 ∞, $g^{t_{k(j)}}_{n(j)}$ converges to g^t. Recall that d^* is the sup metric. It follows from Lemma 3(i) that $g^{t_{k(j)}}_{n(j)}$ converges to g^t in the sup metric d^* as well. Hence, for any $\epsilon > 0$, there exits N such that $k \geq N$, $n \geq N$, $d^*(g^{t_k}_n, g) < \epsilon$. Since each ι_n is continuous, we can choose $k > N$ large enough so that $d(f^k_n, f_n) < \epsilon$ for all $n \leq N$. Hence,

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

proving the continuity of ι. Note that

$$\psi(\iota(t), \iota(s)) = \bigcap_{n \geq 1} f^t_n(\iota(s)) = \lim \Gamma(t, D^s_n) = \Gamma(t, s)$$

Hence Θ is isomorphic to M as desired.

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

$$\iota_n(t) = \iota_n(s) \text{ if and only if } D^t_n = D^s_n \tag{$A1$}$$

Note that this property uniquely identifies the function ι. That is, if $\hat{\iota}$ is any function that also satisfies $(A1)$, $\hat{\iota} = \iota$. To see this note that if $\hat{\iota}_0$ satisfies $(A1)$ then obviously,

$\hat{\iota}_0 = \omega = \iota_0$. Then, a simple inductive step yields the desired conclusion. To see that $\hat{\iota}$ satisfies $(A1)$, note that since it is a isomorphism, we have $\omega = \hat{\iota}_0$ and hence $(A1)$ 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 $f^t_{n+1} = f^s_{n+1}$. Then, the inductive hypothesis yields $D^t_{n+1} = D^s_{n+1}$. Conversely, suppose $D^t_{n+1} = D^s_{n+1}$. 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 $\psi(\hat{\iota}(t), \hat{\iota}(D_n)) = \psi(\hat{\iota}(t), \hat{\iota}(D_n))$ 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 $s \in D^t_n \in \mathcal{D}_n$ for all n. Since ι is an isomorphism, we have

$$f^t_n(\iota(D_n)) = \psi(\iota(t), \iota(D_n)) = \Gamma(t, D_n) = \Gamma(s, D_n) = \psi(\iota(s), \iota(D_n)) = f^s_n(\iota(D_n))$$

for all $n, D_n \in \mathcal{D}_n$. By $(A1)$, we have $\iota(t) = \iota(s)$. Since ι is one-to-one, we conclude $s = t$. This concludes the proof of step 2. □

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. □

7.3 Proof of Theorem 3

Algebras and partitions can be represented as functions: Let $h : S \to X$ be any onto function. This h yields the partition $\mathcal{T}_h = \{h^{-1}(x) \,|\, x \in X\}$ and the algebra $\mathcal{A}_h := \{h^{-1}(V) \,|\, V \subset X\}$. Conversely, for any partition $\mathcal{T}$, the canonical mapping τ of $\mathcal{T}$ (that is, $h : S \to \mathcal{T}$ 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 partition of $\mathcal{T}$ and its canonical mapping τ interchangeably. Note also that for any algebra $\mathcal{A}$, $\tau_{\mathcal{A}}$, the set of minimal elements in $\mathcal{A}\backslash\{\emptyset\}$ is a partition and $\tau_{\mathcal{A}}$ interpreted as the canonical mapping represents $\mathcal{A}$; that is, $\mathcal{A} = \mathcal{A}_{\tau_{\mathcal{A}}}$.

Let $h : S \to X$ and $k : S \to Y$ be two onto functions and define $(h, k)(s) = (h(s), k(s))$ for all s and $h * k : S \to Z$ 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 $A^1, \ldots, A^n \in \mathcal{T}_h * \mathcal{A}_k$ 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 $(h*k)(s) \subset W$ 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 $\{x_2, \ldots, x_n\}$ 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 $m = 2, \ldots, n$. Note that $A^1 = \mathcal{T} * k^{-1}(x)$, $A^m = \mathcal{T} * (S \backslash k^{-1}(x_m))$ for $m > 1$, and $B = \bigcap_{m=1}^n A^m$ as desired. □

For any IPM (T, γ), we can construct an equivalent IPM-EF $\{S, \mathcal{T}_1, \mathcal{T}_2, \nu\}$ 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, γ).

An algebra $\mathcal{A}$ 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 Λ be the lattice of all symmetric algebras. The set $\Lambda_{\mathcal{L}^p} := \{\hat{\mathcal{L}} \in \Lambda \mid \mathcal{L}^p \subset \mathcal{L}\}$ is a sublattice of Λ 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; 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 partition $\mathcal{D}$ induced on T by the function $h_1 \cdot \zeta_1$. By symmetry, this is the same partition as the one induced by $h_2 \cdot \zeta_2$. Since $\mathcal{T}_i * \mathcal{L} \neq \mathcal{A}(\mathcal{T}_i)$, $\mathcal{D} \neq \{\{t\} \,|\, t \in T\}$. Since $\mathcal{L}$ is a fixed point of F, we can assume

$$k = (\nu, h_1 * k, h_2 * k) \tag{$A2$}$$

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

$$k(h_2^{-1}(h_2(t, t^*))) = k(h_2^{-1}(h_2(t', t^*))) \tag{$A3$}$$

It follows from equation $(A2)$ above that $D^{t^*} \neq D^{t''}$ implies $k(t, t^*) \neq k(t', t'')$. Therefore, equation $(A3)$ implies

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

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, (T, γ) is not valid .

For the converse, let $\mathcal{D}$ be a partition other than the finest that satisfies (ii) in the definition of validity. Define $h_1(t, t^*) = \{(t', t'') \,|\, t' \in D^t\}$ and $h_2(t^*, t) = \{(t'', t') \,|\, t' \in D^t\}$. 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. □

7.4 Proof of Theorem 4

For any ordered IPM (T, δ), let $\Delta(t, t') = (\delta(t, t'), \delta(t', t))$. Then, the definition validity for an ordered IPM replaces Γ with Δ.

Lemma 9: *Let $M = (T, \delta)$ be a simple IPM and suppose δ is nondecreasing in both arguments. Then, M is a reciprocity model if and only if $\delta(x, \cdot) = \delta(z, \cdot)$ implies $x = y$.*

Proof: Let $M = (T, \delta)$ be a simple IPM and assume δ is nondecreasing in both argument but $\delta(x, \cdot) = \delta(z, \cdot)$ for some $x \neq y$. Then, define the partition $\mathcal{D}$ as follows: for $y \notin \{x, z\}$, $D^y = \{y\}$, and $D^x = D^z = \{x, z\}$. It follows that $\delta(x, \cdot) = \delta(z, \cdot)$ and therefore $\delta(\cdot, x) = \delta(\cdot, z)$ and hence $\Delta(w, D) = \Delta(w', D)$ for all $w \in T$, $w' \in D^w$, and $D \in \mathcal{D}$. Hence, M is not valid.

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

$$\Delta(w, \bar{D}_2) = \Delta(w', \bar{D}_2) \qquad (A4)$$

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 $\Delta(w_n, y_n) = \Delta(w'_n, y'_n)$ 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 Δ ensures $\Delta(w, y) = \Delta(w', y')$ and proves $(A4)$.

The weak monotonicity of δ in both arguments together with $(A4)$ implies

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

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

Lemma 9 establishes that any simple increasing IPM is a reciprocity model. Next, suppose $M = (T, \delta)$ is a reciprocity model. By definition, $\succeq$, the nicer than relation is transitive and since M is reciprocity model it is also complete. The continuity of γ yields the continuity of $\succeq$. Since T is a compact metric space, it is separable and hence there exists a continuous real-valued function $x : T \to I\!R$ that represents $\succeq$. Let $K := x(T) = \{x(t) \,|\, t \in T\}$. Let $D^t = \{t' \in T \,|\, x(t') = x(t)\}$ and $\mathcal{D} = \{D^t \,|\, t \in T\}$. Clearly, $\mathcal{D}$ is a partition of T such that $\delta(t', \cdot) = \delta(t, \cdot)$ for all $t' \in D^t$. Since every type reciprocates, $\delta(\cdot, t') = \delta(\cdot, t)$ for all $t' \in D^t$ and therefore $\Delta(t, D) = \Delta(t', D)$ for all $t \in T$, $t' \in D^t$ and $D \in d$. 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, $\delta(w,y) = \delta(x^{-1}(w), x^{-1}(y))$ for all $w, y \in K$. Since x and δ are continuous, so is δ'. It follows that (K, δ') is isomorphic to M and therefore is valid. Since x represent $\succeq$ and every type in M reciprocates, δ' is weakly increasing in both arguments. Finally, Lemma 9 and the validity of M imply (K, δ') is increasing. □

7.5 Proof of Theorem 5

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

Then, without loss of generality, assume $T' = \{1, \ldots, m\}$. 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 $T' = \{1, \ldots, k+1\}$. Validity ensures that $\delta'(t, \cdot)$ 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 $T = T' \backslash \{t\}$ and let δ be the restriction of δ' to $T \times T$.

First, we argue that (T, δ) must be a valid binary symmetric IPM. If not, there exist a some partition other than the finest that satisfies (ii) in the definition of validity. Then, let $\mathcal{D}' = \mathcal{D} \cup \{\{\mathcal{T}\}\}$ and note that $\mathcal{D}'$ proves that (T', δ') is not valid, a contradiction. Hence, by the inductive hypothesis, we can assume that $T = \{1, \ldots, k\}$ and either $\delta = \delta_k$ or $\delta = \delta_{k+1}$. In the former case, we have $\delta'(k, t') = \delta(k+1, t') = 1$ for all $t' \in T'$, contradicting the validity of (T', δ'). In the latter case, we have $(T', \delta') = M_{k+1}^0$ as desired. □

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. Dufwenberg (2009) “Dynamic Psychological Games,” *Journal of Economic Theory,* 144(1), 1–35.

Battigalli, P. and M. Siniscalchi (2003): “Rationalization and Incomplete Information,” *The B.E. Journal of Theoretical Economics*, 3(1).

Berg, J., Dickhaut, J. and K. McCabe (1995) “Trust, Reciprocity and Social History,” *Games and Economic Behaviour*, 10, 122–42.

Bergemann, D. and S. Morris (2007) “Strategic Distinguishability with an Application to Virtual Robust Implementation,” mimeo, Princeton University.

Bergemann, D. and S. Morris (2009) “Robust Virtual Implementation,” *Theoretical Economics,* 4(1), 45–88.

Bergemann, D. and S. Morris and S. Takahashi (2010) “Interdependent Preferences and Strategic Distinguishability,” mimeo, Princeton 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.

Brown, A. and C. Pearcy (1995) *An Introduction to Analysis,* Springer-Verlag, New York.

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,” *Economics Letters,* 12(2), 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 (2007) “A Tractable Model of Reciprocity and Fairness,” *Games and Economic Behavior,* 59(1), 17–45.

Dal Bo, P. (2005) “Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games,” *American Economic Review*, 95(5), 1591–1604.

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

Dekel, E., D. Fudenberg and S. Morris (2006b) “Interim Rationalizability,” *Theoretical Economics*, 2, 15–40.

Dohmen, T., Falk, A., Huffman, D. and U. Sunde (2009) “Homo Reciprocans: Survey Evidence on Behavioral Outcome,” *The Economic Journal,* 119, 592–612.

Dufwenberg, M. and G. Kirchsteiger (2004) “A Theory of Sequential Reciprocity,” *Games and Economic Behavior*, 47, 268–98.

Ely, J. and M. Peski (2006) “Hierarchies of Beliefs and Interim Rationalizability,” *Theoretical Economics*, 1, 19–65.

Falk A., E. Fehr and U. Fishbacher (2008) “Testing Theories of Fairness - Intentions Matter”, *Games and Economic Behavior,* 62(1), 287–303.

Falk A. and U. Fishbacher (2006) “A Theory of Reciprocity,” *Games and Economic Behavior,* 54(2), 293–315.

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 (1989) “Psychological Games and Sequential Rationality,” *Games and Economic Behavior,* 1, 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.

Mariotti, T., Meier, M and M. Piccione (2005) “Hierarchies of Beliefs for Compact Possibility Models,” *Journal of Mathematical Economics,* 41(3), 303–324.

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 (2007) “Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings,” *Journal of Economic Theory,* 136(1), 197–216.

Sobel, J. (2005) “Interdependent Preferences and Reciprocity,” *Journal of Economic Literature,* 43(2), 392–436.