Chapter 3

INTERPRETATION, COORDINATION AND CONFORMITY

Hykel Hosni
Scuola Normale Superiore, Pisa
hykel.hosni@sns.it

Abstract The aim of this paper is to investigate a very general problem of (radical) interpretation in terms of a simple coordination game: the *conformity game*. We show how, within our mathematical framework, the solution concept for the conformity game does indeed provide an algorithmic procedure facilitating *triangulation*, in the sense of Davidson.

3.1 Introduction

Suppose that the robotic rovers *I* and *II* are conducting a joint operation on a terrain about which nothing was known to their designer (say the units are operating on Mars). Suppose further that communication among the units has been lost and that the only way *I* and *II* have to restore it is to meet on some location $\mathtt{l}$, chosen from a finite set of possibilities equally accessible to both. Assuming that any location is as good as any other, provided that *I* and *II* agree on it, how could the robots reason and act so as to facilitate their meeting? That is, how should they *choose* $\mathtt{l}$?

We see situations of this sort as instantiations of interpretation problems. After all, what *I* and *II* must do in order to restore communication is to (i) attach a certain meaning to the representation they have of their environment, (ii) form expectations about each other's behaviour, and (iii) act accordingly. More specifically, once the possible locations, say $\mathtt{l}_1, \ldots, \mathtt{l}_k$, are identified, given their common intention, agents must interpret each other relative to the 'external world'—the environment in which they happen to operate—so as to increase their chances of agreeing on the final choice of a location. Since *I* and *II* do not share a language, in fact they cannot communicate, the problem they face is one of *radical* interpretation.

At the same time, this situation is a clear example of strategic interaction: what corresponds to the 'rational' or 'commonsensical' or even 'logical' or

O. Majer et al. (eds.), *Games: Unifying Logic, Language, and Philosophy,*
Logic, Epistemology, and the Unity of Science 15,

simply 'best' course of action for *I* depends on the course of action adopted by *II* (and the other way round). This quite naturally suggests that game theory might somehow provide us with precise and well-understood guidelines for the mathematical solution of our problem. As will be shortly illustrated, however, for the kind of strategic interaction that we shall be concerned with, the classical solution concepts studied in the theory of non-cooperative games are of no use whatsoever.

The framework of Rationality-as-Conformity, recently introduced by Jeff Paris and the present author (see Hosni and Paris, 2005; Hosni, 2005), attempts to define, within an abstract mathematical setting, 'rationality' in situations of strategic interaction of the sort mentioned above. It is the purpose of this paper to illustrate how such a mathematical characterization of rationality can be used to provide a solution concept for problems of (radical) interpretation, whenever the latter is considered in terms of games of (pure) coordination.

The paper is organized as follows. First (Sections 3.1.2–3.1.4), we isolate the fundamental aspects of radical interpretation problems in connection with the interactive choice problem considered in the Rationality-as-Conformity framework. Putting forward the intrinsic strategic nature of the problem of radical interpretation leads us to formulate it mathematically in terms of the *conformity game*, fully described in Section 3.2. Being a game of multiple (indiscernible) Nash-equilibria, the conformity game is indeed a (pure) *coordination game* and as such, it is generally regarded to be unsolvable within the traditional game-theoretical framework of non-cooperative games. We discuss in Section 3.2.1 the informal constraints that an adequate solution concept for the conformity game should satisfy and move on towards formalising the solution concept for the conformity game in Section 3.3. This is based on the Minimum Ambiguity Reason, introduced in Honsi and Paris (2005) as part of the Rationality-as-Conformity framework. We will then conclude by showing that this solution concept indeed provides an algorithmic solution for establishing communication—triangulating—in problems of radical interpretation.

Radical interpretation helps in clarifying the issues and the assumptions underlying a basic characterization of 'rationality' in communicationless scenarios yet without immediately providing any effective procedure to achieve it. Pure coordination games, on the other hand, help framing a variety of possible solution concepts based on *saliency*, which however seem to lack of a general formal structure allowing us to evaluate their 'rational' underpinnings. This paper attempts to unify the fundamental aspects of both frameworks by means of the mathematical abstraction provided by Rationality-as-Conformity.

Many connections between (linguistic) interpretation and (coordination) games have been explored, from the classic investigation by Lewis (1969) to the game-theoretic accounts of linguistic interpretation of Parikh (2000) and van Rooy (2004). Though Lewis considers the 'use of language' as a particular

kind of 'coordination problem' (Lewis, 1969), the present author has no knowledge of any attempt to relate mathematically the structure of *pure coordination* games with that of *radical* interpretation.

3.1.1 Why rationality-as-conformity?

As illustrated at length in Honsi and Paris (2005), we understand 'conformity' as the adoption of a choice process facilitating the selection of the same possible world (say a location 1 in the robotic rover example above) as another like-minded yet otherwise inaccessible agent.

Within frameworks of this sort, solid arguments can be put forward supporting the view that commonsensical agents not only happen to be generally able to conform, they should indeed aim at conforming if they are to be rational.

1. *The members of a society have a natural inclination to coordinate successfully*. This is a conclusion of the numerous empirical investigations that have been carried out during the last decades in the area of *behavioural game theory*, following Schelling's early intuitions about *coordination games* (Schelling, 1960) (see e.g. Mehta et al., 1994; Camerer, 2003). The common pattern of those investigations puts forward that, whenever, say, pairs of agents face a strategic choice problem in which they have a joint motivation (intention) to coordinate their solutions, they will be able to adopt certain kinds of choice processes facilitating this coordination. In other words, there are reasons to believe that principles, strategies and patterns of choice behaviour exist which, if adhered to, will result in agents having generally better chances to coordinate (and never strictly worse) as they would have, should they adopt random patterns of behaviour.

2. *Agents satisfying probabilistic 'commonsense' should end up assigning similar degrees of belief*. This is a consequence of a number of a contributions in the area of subjective probability logic. In the normative framework developed by Paris and Vencovská (1990, 2001) and Paris (1994) a small number of so-called commonsense principles are identified and it is shown that, if adhered to, those principles uniquely and completely determine any further assignment of probabilities, i.e. degrees of belief. This distribution of probabilities, the one with the largest possible entropy, is provably the only one jointly consistent with the (probabilistic) knowledge possessed by an agent and those principles. Hence, similar agents, possessing similar knowledge bases and applying the inference process identified with commonsense, all assign similar degrees of belief to the as yet undecided sentences.

3. *'Rationality is a social trait. Only communicators have it.'* This is the conclusion of Davidson (2001). The idea here is that a necessary condition for rationality is an adequate apparatus for communication, which in turn requires agents to be able to move from a condition of mutual inaccessibility (no shared language), to a condition in which communication is being enabled. This transition implies that agents are attaching similar meanings to the publicly accessible causes of their reciprocal choice behaviour. This aspect of Rationality-as-Conformity, which Donald Davidson calls *triangulation*, and its underlying structure are the main topic at focus in the rest of this paper.

3.1.2 Radical translation and the Principle of Charity

Put roughly, a problem of *radical translation* is one in which one agent—a linguist in the field—is trying to build up a 'translation manual' accounting for the utterances of a native speaker of a language about which the linguist has no knowledge whatsoever. This complete lack of information, together with the fact that the two agents are assumed not to share a third language, make the translation problem *radical*.

The radicalness of the situation induces Quine to observe that a hypothetical theory of radical translation should start by relating the native's linguistic behaviour to the one the translator would adopt, were she to be in the same 'observable situation' as the native.

In his classic example Quine, who was the first to introduce this problem in connection with the translation of logical constants (Quine, 1960, 2), imagines that the native speaker utters the expression GAVAGAI in correspondence of a rabbit passing by, causing—possibly on repetitions of similar events—the translator to conjecture that GAVAGAI translates into 'rabbit'.

There are many subtleties connected with this example, none of which being of particular interest for present purposes. Rather, two issues involved in the radical translation exercise are relevant for our present discussion:

1. What is it, if anything, that *justifies* (epistemologically) the translator in the above conjecture?

2. How far can the translator go in relying on this conjecture?

Those questions are clearly not unrelated. The former calls for the observation that a linguist may just introspect and conclude that "as a native speaker of English, I would utter RABBIT were that kind of animal to pass by". This subjunctive is clearly grounded on the assumption that the linguist and the native speaker, though lacking of a shared language, are nonetheless *like-minded* individuals and hence inclined to adopt similar linguistic behaviours under similar

(observable or conceivable) circumstances. Elevated to the status of a normative maxim, this is known as the *Principle of Charity*.

Any reasonable understanding of this principle, of course, asks for a clarification of what is meant by 'similar linguistic behaviour' as well as 'similar observable (conceivable) circumstances' and in the natural language case these are by no means trivial clarifications to do and many criticisms to the adoption of the principle seem to pivot on this difficulty (see e.g. Feldman, 1998; Wachbroit, 1987; McGinn, 1977 for the role of the principle in the explanation of rationality, and Nozick, 1993, 152–158; Glock, 2003, 194–199 for more forceful criticisms). It turns out, however, that in the abstract and simplified mathematical framework of Rationality-as-Conformity, correlated notions can be defined rigorously and put to work in the formal characterisation of rational choice behaviour in the absence of communication or learnt conventions.

The second crucial feature of radical translation problems relates to their fundamental indeterminacy. Quine argues that there cannot be a unique translation manual which the linguist in the field may be able to construct. Rather, there must be a plurality of manuals, all equally acceptable, that is to say, equally supported by the available evidence. The only attempt that the linguist can do to reduce this indeterminacy is the application of the Principle of Charity, leading her to *discard* all those possible translation choices that will make the native utterances systematically wrong (or incoherent), by the translator's lights. After this 'rational' refinement, the choice of a translation manual may simply be underdetermined by the empirical evidence available to the translator.

That 'rationality' might not always lead to a unique choice (without randomisation) is a feature captured by the Rationality-of-Conformity framework as well. Indeed, some problems might just be too hard to admit of a unique solution.

3.1.3 Radical interpretation and triangulation

The issues of radical translation and charity are taken a step further by Davidson's investigations on *radical interpretation*. For the purposes of the present discussion, the main points of departure of the situation described in the radical interpretation problem with respect to the one discussed in connection with radical translation can be outlined as follows. Davidson does not assume that agents are native speakers of distinct languages. He rather assumes that they do not have a shared language whatsoever and that their goal consists in establishing communication.

The Principle of Charity is thus sharpened and indeed assumed to be a necessary condition for the manifestation of rational behaviour tout court. Moreover, the interpretation problem is grounded on a fundamental symmetry which need not hold in the translation case, that is that both agents share a common

intention to communicate: the interpreter wants to understand the interpretee who, in turn, wants to be understood by the interpreter.

Differences in the formulation of the problem lead to differences in the proposed solutions. Quine's major problem is that of locating the common cause of the linguistic behaviour, which he identifies in the 'stimulus-meaning'. Davidson overcomes many of the difficulties related to this concept by introducing the metaphor of *triangulation*. While Davidson takes charity as a presumption of rationality upon which the possibility of interpretation and mutual understanding themselves rest, he acknowledges that it can only provide a 'negative' contribution, namely by guiding the interpreter towards *discarding* possible interpretations which would systematically make the interpretee wrong or incoherent to her own lights. Triangulation, on the other hand, is the recognition that the similarities observed in each other's linguistic behaviour find their common cause in the same portion of the external environment shared by the agents. It is the location of those causes that results in getting a clue about the other's meanings.

Davidson introduces triangulation by considering a 'primitive learning situation', in which a child learns to associate the expression "table" to the actual presence of a table in a room. The way the child can learn to do so, relies in her ability to generalise, to discover and exploit similarities among situations. Sharing similar generalisation patterns is what makes the child's response to the presence of a table—the utterance of the word "table"—meaningful to us. This is the rational structure that agents must have in order for communication to start.

> The child finds tables similar; we find tables similar; and we find the child's responses in the presence of tables similar. It now makes sense for us to call the responses of the child responses to tables. Given these three patterns of response we can assign a location to the stimuli that elicit the child's responses. The relevant stimuli are the objects or events we naturally find similar (tables) which are correlated with responses of the child we find similar. It is a form of triangulation: one line goes from the child in the direction of the table, one line goes from us in the direction of the table, and the third line goes between us and the child. Where the lines from child to table and us to table converge, 'the' stimulus is located. Given our view of child and world, we can pick out 'the' cause of the child's responses. It is the common cause of our response and the child's response. (Davidson, 2001, 119)

A fundamental aspect of the triangulation process, then, consists in the recognition of the role played by constraints imposed by the 'external world' on the interpretational choices. In particular, the interpreter should ascribe 'obvious beliefs' (e.g., the presence of a table) to the interpretee, and project onto her the likewise 'obvious' consequences (that she will behave accordingly). Suppose, for instance, that rover *I* in the initial example perceives the presence of a perfectly round crater. According to this way of reasoning, *I* should expect *II*

to be able to perceive the crater as a perfectly round one. At the same time *II* should expect *I* to expect that *II* itself would perceive the crater as a perfectly round one etc., and of course consider this as a relevant feature for the selection of the rendez-vous location 1. This 'like-mindedness' or 'common reasoning' of agents plays a fundamental role in the Rationality-as-Conformity framework and constitutes the main conceptual fulcrum on which the present analysis of interpretation, coordination and conformity pivots.

As for translation, in the case of interpreting natural language triangulation presents several difficulties mostly related to the rigorous explanation of what intervenes in the 'recognition of the common causes' of common linguistic behaviour. A recent comprehensive discussion on the topic can be found in Glock (2003). What is relevant for us here, however, is that the complication of considering the full case of interpreting natural language is surely one of the reasons why the theory of radical interpretation does not seem to allow for a clear-cut *procedure* by means of which agents can achieve, or at least facilitate, triangulation.

Within the mathematical framework of Rationality-as-Conformity we are able to provide one such effective procedure. It goes without saying that the structure therein considered (comparable to unary predicate languages) is much weaker than the one required by Davidson for the construction of a theory of meaning, namely the full first-order logic with equality. Our hope is, of course, that of eventually extending the results obtained in this initial framework to cover more 'realistic' situations.

3.1.4 Radical interpretation as coordination

Thomas Schelling is usually credited with the introduction of *coordination* problems in the game-theoretical literature. Roughly speaking, a tacit coordination game is a situation of interdependent, strategic choice characterised by the absence of communication among players who nonetheless aim at performing the same choice—i.e., coordinating. Schelling's example concerns a couple who get accidentally separated in a supermarket and want to rejoin.

Schelling calls this a problem of 'tacit coordination' with 'common interests' and notices that given the lack of communication—which indeed makes the coordination *tacit*—all that agents can rely on are the assumption of like-mindedness and the mutual expectations that this generates. What Schelling intends to discuss is the characterisation of 'rational rules' accounting for the ability humans have to coordinate in the complete absence of communication.

The situation described by Schelling is one of radical interpretation for which a triangulation-like solution is advocated. Indeed, after introducing the supermarket problem he goes on commenting as follows:

> What is necessary is to coordinate predictions, to read the same message in the common situation, to identify the one course of action that their expectation of

> each other can converge on. They must 'mutually recognize' some unique signal that coordinates their expectations of each other. We cannot be sure that they will meet, nor would all couples read the same signal; but the chances are certainly a great deal better than if they pursued a random course of search. (Schelling, 1960, 54)

The analogies with the solution proposed by Davidson for the radical interpretation problem stand out: both charity and triangulation appear clearly in Schelling's illustration of the fundamental features of the solution concepts adequate for tacit coordination games. Entirely analogous remarks can be made in relation to 'tacit agreement' as discussed by Lewis in his classic work on conventions (Lewis, 1969).

3.1.5 Towards a solution concept

What facilitates conformity in coordination problems of the sort introduced above is, according to the investigations initiated by Schelling, the selection of those possible options—strategies—that would be perceived by agents as *focal points*. Indeed, the many investigations that followed Schelling's original intuitions can be seen as attempts at providing an explanation for the ability that human agents have in exploiting focal points for the purpose of coordinating.

The intuition underlying the use of focal points is that these correspond to strategies which enjoy some degree of 'saliency' or 'conspicuousness', in Schelling's phraseology, which will lead agents to in fact focus on certain options instead of others. Distinctions are made then, on what saliency can be taken to be (see, e.g. Sugden, 1995; Kraus et al., 2000). For present purposes we will concentrate on salience as given by the identification of a *choice process* which an agent might adopt upon reflection about which choice process another like-minded agent with a common intention to coordinate might herself adopt. In the literature this is usually referred to as *Schelling's salience*.

The most distinctive feature of salience is the combination of *uniqueness* and *obviousness* of focal points. These are thought of as options which somehow *stand out* when considered in the context of the strategies available to the agents in a given coordination problem. So, for example, the robotic rovers of our initial example will base their choice on saliency if they will select a location $\mathtt{l}$ which stands out in the set $\{\mathtt{l}_1, \ldots, \mathtt{l}_k\}$. Naturally, if I can conclude that the location $\mathtt{l}_j$ does indeed stand out, the fact that II intends to conform to the choice it expects I to make will lead, together with the assumption that I and II are like-minded, to the conclusion that $\mathtt{l}_j$ is the *obvious* choice for this problem.

It is in this spirit that Schelling suggests that, in order for agents to coordinate successfully, they must 'mutually recognize a unique signal'. Intuitive as it may be, however, a lighthearted resort to 'uniqueness' can prove to be rather tricky. As it has been put forward by (Kraus et al., 2000), this becomes

a major concern once we take into account the limitations (i.e., bounded reasoning capabilities) of the agents. Moreover, there could be circumstances in which appeal to uniqueness may lead to undesirable conclusions, as we will have occasion to notice below.

In what follows, we will rather attempt at formalizing the notion of a focal point by characterising saliency in terms of the *minimisation of the ambiguity* of the options available to the agents. In order to do this we shall firstly provide a mathematical formalisation of the *context* within which focal points are to be discerned. This will enable us to study the corresponding *reasoning process*, that is to say an algorithm for the determination of the minimally ambiguous strategies within the context.

3.2 The conformity game

In the spirit of the Rationality-as-Conformity approach, we tackle the knowledge representation issue by considering the simple model in which options are the *possible worlds* generated by mapping a finite set A to the binary set $2 = \{0, 1\}$. Nothing else is assumed about the structure of the set A.

The domain of the game is $\wp^+(2^A)$, the set of non-empty subsets of 2^A which denotes the set of all possible worlds. We attach to elements $K \in \wp^+(2^A)$ an epistemic value, namely we take players to have common knowledge of the fact that the options they have to choose from are those in K, which includes the possible world which will be eventually selected. Intuitively, then, the the cardinality of K gives a quantitative measure of the agents' uncertainty about the other's actual choice.

The *conformity game* is a two-person, non-cooperative game whose normal form goes like this: Each player is to choose one strategy out of a set of possible choices, identical for both agents up to permutations of A and 2, where each strategy corresponds to one element of $K = \{s_1, \ldots, s_k\}$, say. Strategies are therefore represented in this game as finite binary strings. Players get a positive payoff p if they play the same strategy, and nothing otherwise, all this being common knowledge. (Figure 3.1 represents the conformity game for $k = 3$.)

		Player *II*		
		s_1	s_2	s_3
Player *I*	s_1	p, p	$0, 0$	$0, 0$
	s_2	$0, 0$	p, p	$0, 0$
	s_3	$0, 0$	$0, 0$	p, p

Figure 3.1: The conformity game

Note that, for present purposes, we limit ourselves to the case in which each identical pair of strategies yields a unique positive payoff p, so that any point in

the diagonal would be as good as any other as far as the agents are concerned: all that matters is that they conform on their world-view.

Being a game of multiple Nash-equilibria in which the players are assumed to be inaccessible to each other, the conformity game is a typical example of a (pure) *coordination game*, a kind of game which is generally considered to be unsolvable within the traditional theory of non-cooperative games. (See, e.g., Camerer, 2003 for a discussion on coordination problems other than 'pure'.)

Before going into any further details of the conformity game it will be useful to introduce some ideas concerning the selection of multiple Nash-equilibria in pure coordination games, and relate these to the intuitions underlying the conformity game.

3.2.1 Multiple Nash-equilibria and the conformity game

Traditional game theoretic solution concepts usually characterize distinguishability among options (strategies) in terms of the comparison of (ordinal) utilities, 'rationality' being defined in terms of utility maximization. As an immediate consequence of this, whenever options are perceived by an agent as being equally desirable—i.e., payoff-indistinguishable—the selection of strategies usually referred to as 'rational' turns out to be unhelpful as solution concept.

Here is where the concept of 'rationality' pursued in the Rationality-as-Conformity framework shows its most relevant point of departure from the game theoretic tradition. In the former, in fact, rationality is not defined in terms of maximisation of utility, but on the mutual expectations of agents sharing a common intention. Hence the conformity game is characterized by a complete symmetry with respect to both payoffs and players. Moreover, the possibility of considering 'extra structure' in the game by focusing on its presentation can be ruled out by means of appropriate mathematical devices, to be shortly introduced. Hence, in Schelling's terminology, the conformity game is a 'clueless', 'genius-proof' game.

To appreciate the point further, recall that the typical solution concept for non-cooperative games introduces a notion of distinguishability among strategy profiles—Nash-equilibrium—which is in fact weaker than simple pay-off dominance. If a Nash-equilibrium exists, yet is not unique, then a natural way of reducing the situation to the standard case would just involve selecting the equilibrium, if one exists, with the the highest possible payoff. In particular, it can happen that a strategic game admits of say two equilibria with distinct ordinal utilities, which nonetheless are, according to the theory of Nash-equilibrium, undistinguishable. Due to its wide applicability, a largely studied example is the following variant of the game known in the literature as the

Battle of the sexes (see, e.g., Osborne, 2004). Two players are to choose between a pair of options for a night at the concert hall (say, B and S, for Bach and Stravinsky) with the distinctive feature that whilst both players strictly prefer the same option (say B), they are still entitled to choose (S, S), a Nash-equilibrium of this game. The idea here being that although they both prefer going to the Bach concert, they still prefer going to the Stravinsky concert together rather than going to different concerts. In games of this sort, the theory of Nash-equilibrium gives agents exactly the same reasons for playing a payoff-dominated strategy as for playing a payoff-dominant one.

The conformity game, as any pure coordination game, pushes this limitation of the theory of Nash-equilibrium even further, given that the obvious refinement which would lead agents to select, among the Nash-equilibria, the one with the highest payoff (if this exists), cannot be applied due to the complete symmetry of the payoffs. Similar considerations apply to risk-dominance, the 'cautious' dual of payoff-dominance (Harsanyi and Selten, 1988).

It follows that traditional solution concepts are generally inadequate for the conformity game, and indeed for any other game of (pure) coordination. The general feeling on the matter can be illustrated by recalling Schelling's own words (1960):

> Poets might do better than logicians at this game, which is perhaps more like 'puns and anagrams' than like chess. (Schelling, 1960, 58)

An entirely similar attitude is shared (4 decades later) by Camerer, who indeed argues in favour of the empirical (behavioural) investigation on the way players choose among equilibria. As to the 'logical' approach, he remarks that

> [t]his *selection* problem is unsolved by analytical theory and will only be solved by observation. (Camerer, 2003)

Still, as noted by Schelling, players can generally do better than plain randomization in pure coordination games. The extensive empirical investigations that took place over the past decades (see, e.g., Mehta et al., 1994; Sugden, 1995; Janssen, 1998, as well as the results of computer simulations Kraus et al., 2000, strongly support Schelling's early insight that there are in fact choice processes that can facilitate conformity [i.e., that lead agents to coordinate their choice better than plain randomization]).

In the remainder of this paper we will provide a formalisation of a solution concept for the conformity game which is based on the considerations about salience and is underpinned by the principle of charity discussed in Section 3.1.3.

3.3 Solving the conformity game

Recall that the key element intervening in the representation of the conformity game is given by possible worlds, which in the present interpretation amount to the strategies available to the players. We clearly have two

possibilities: either worlds (strategies) in K have no structure other than being distinct elements of a set, or worlds in K do have some structure and in particular there are properties that might hold (be true) in (of) some worlds. In the former case we seem to be forced to accept that agents have no better way of playing the conformity game other than picking some world $f_i \in K$ at random (i.e., according to the uniform distribution). In the latter case, however, agents might use the information about the structure of the worlds in K to focus on some particularly 'distinguished' option to be taken as a focal point.

Consider, for example, the simple case in which worlds (strategies) are maps $f : 4 \longrightarrow 2$ and suppose $K = \{f_1, f_2, f_3, f_4, f_5\} \subseteq 2^4$ is presented as the matrix in Figure 3.2.

	0	1	2	3
f_1	0	0	0	1
f_2	0	1	0	0
f_3	0	1	1	0
f_4	1	1	1	1
f_5	0	0	1	0

Figure 3.2: A representation of the strategy set K

We know from the strategic representation of the conformity game that each pair of identical strategies yields the same utility, so players who intend to conform must look for salient properties to characterize some of the options as those which are likely to be selected by another agent. At the same time, however, we want to rule out the possibility that agents will take into account inessential properties of the set K as being salient, so our first goal is that of ensuring the complete symmetry of the representation. A way of achieving this consists in informing each agent that it is being presented with a matrix K (for instance the one illustrated in (2) which agrees to the one faced by the other player only up to permutations of A and permutations of 2, that is to say, only up to permutations of the columns (and of course rows) of the matrix as well as the uniform transposition of 0's and 1's.

On the assumption of like-mindedness, i.e. common reasoning, if one of those binary strings, say f_j should stand out as having some distinguished properties, agents will conclude that such properties are indeed intersubjectively accessible and hence select f_j. In this way players will go about producing a *reason* for selecting the option f_j. We now move on to formalize this notion.

3.3.1 Introducing asymmetries with reasons

Given the inapplicability of the payoff-dominance principle to the conformity game, the analogy with coordination games suggests that in order to facilitate triangulation we need to introduce some *asymmetries* among the strategies available to the players of the conformity game. We propose here

to formalise this by means of a choice process derived from the *Minimum Ambiguity Reason* introduced in Hosni and Paris (2005).

In a nutshell, the construction of this choice process, or Reason, takes place by means of identifying certain selection principles that players of the conformity game might come to tacitly agree upon, given the goal of the game and their common knowledge of it. This construction will adhere to the charity principle recalled above, in that it is pivoted on the idea that the only clue available to the players about each others' world view is that they share common reasoning.

We define a Reason R to be a choice function from the domain of the conformity game $\wp^+(2^A)$ to itself such that $R(K) \subseteq K$. The general intuition, as discussed in connection with radical interpretation, is that agents should apply Reasons to discard those possible strategies that will prevent them from conforming on their mutual expectations. Given the like-mindedness assumption and the fact that the size of K is proportional to the uncertainty of the players about each other's behaviour, it can be immediately appreciated that a *perfect reason* will be a choice function which always returns a singleton, a unique strategy. It is likewise immediate to see, however, that we cannot expect this to happen in general. As we learnt from radical translation and interpretation, there can be real indeterminacy in the choice problem at hand.

Hence, if after applying their Reason players are left with a plurality of strategies, they will conclude that the choice problem at hand is just underdetermined with respect to the information they possess (the structure of their binary matrix) and will go about to select at random from $R(K)$. In the worst possible case agents will find that $R(K) = K$. At this opposite extreme from the perfect reason, agents will just realize that the strategies from which the choice is to be made are—to their lights—absolutely undistinguishable.

The construction of the Minimum Ambiguity Reason, then, just amounts to constraining the choice process R in such a way as to facilitate the identification of focal points in the conformity game. This characterization will be provided by means of an effective procedure.

3.3.2 The minimum ambiguity reason

Our first goal is constraining R in a way that will provide an adequate formalisation of the symmetries among the players and the possible strategies. This will lead us to formulate the first requirement imposed on the algorithm for computing $R(K)$, namely that if f and g are, as elements of K, *indistinguishable*, then $R(K)$ should not contain one of them, f, say, without also containing the other, g. In other words, an agent should not give positive probability to picking one of them but zero probability to picking the other. The argument for this is that if they are 'indistinguishable' on the basis of K then another

agent could just as well be making a choice of $R(K)$ which included g but not f. Since agents are trying to make the same ultimate choice of element of K, taking that route may be worse, and will never be better, than avoiding it. Indeed, this requirement can be further motivated by direct reference to the radical interpretation problem. The ideal goal of translation as well as interpretation, consists in individuating systematically synonymy among linguistic expressions. In our abstract mathematical setting, synonymy can be understood as "undistinguishability" among possible worlds. It, therefore, follows that accepting in $R(K)$ only one of a pair of undistinguishable worlds amounts to admitting the systematic violation of synonymy, a most undesirable situation for any theory of interpretation.

The second requirement is that the players' choice of $R(K)$ should be as small as possible (in order to maximize the probability of randomly picking the same element as another agent) subject to the additional restriction that this way of thinking should not equally permit another like-minded agent (so also, globally, satisfying the first requirement) to make a different choice, since in that case any advantage of picking from the small set is lost.

The first consequence of this is that initially the agent should be looking to choose from those minimal subsets of K closed under indistinguishability, 'minimal' here in the sense that they do not have any proper non-empty subset closed under indistinguishability. Clearly, if this set has a unique smallest element then the elements of this set are the least ambiguous, most outstanding, in K and this would be a natural choice for $R(K)$. However, if there are two or more potential choices $X_1, X_2, \ldots, X_k$ at this stage with the same number of elements then the choice of one of these would be open to the obvious criticism that another 'like-minded agent' could make a different (in this case disjoint) choice. Faced with this revelation our agent would realise that the 'smallest' way open to reconcile these alternatives is to now permit $X_1 \cup X_2 \cup \cdots \cup X_k$ as a potential choice whilst dropping $X_1, X_2, \ldots, X_k$.

The agent now looks again for a smallest element from the current set of potential choices and carries on arguing and introspecting in this way until eventually at some stage a unique choice presents itself. We will understand this unique choice as the required focal point, the center of agents' triangulation.

In what follows, we shall give a formalisation of this procedure. All the results to follow have appeared (or are straightforward generalisations of those spelled out) in Hosni and Paris (2005) and Hosni (2005) and therefore the proofs are omitted here.

3.3.3 Transformations

We begin by formalising the intended notion of undistinguishability among worlds in K. In the current abstract mathematical framework this amounts to

providing a formalisation of synonymy among possible options—with respect to the radical interpretation problem—as well introducing a utility-free evaluation (pairwise comparison) of the strategies available to the agents in the conformity game.

The central concept is that of a *transformation* of possible worlds. The intuition to be formalised being that a transformation can act on a set of possible worlds by operating changes that agents should consider inessential to the choice problem they are facing. Hence the possibility of transforming (formally) one world into another one will lead agents to consider these to be indistiguishable.

We define a function $j : K \to 2^A$ a *transformation of* K if there is a permutation σ of A and a permutation δ of $\{0, 1\}$ such that $j(f) = \delta f \sigma$ for all $f \in K$. We shall say that a transformation j of K is *a transformation of* K *to itself* if $j(K) = K$.

The intuition here is that a transformation j of K to itself produces a copy of K—$j(K)$—in which the 'essential structure' of K is being preserved. To see this in practice, simply take the matrix introduced above in Section 3.3, from which the explicit mention of the set A and the labels of the binary strings are omitted, as illustrated in Figure 3.3:

$$\begin{array}{cccc} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{array}$$

Figure 3.3: The matrix representing K

It can be easily seen that putting δ to be the identity function (id) and $\sigma = (1, 2)$ (the permutation transposing 1 and 2 in $\{0, 1, 2, 3\}$), we will obtain the transformation transposing the 'second' and 'third' column of the above matrix. Furthermore, by letting $\sigma' = id$ and $\delta' = (0, 1)$ we obtain a matrix with 0's and 1's exchanged. These can be represented as:

$$\begin{array}{cccc} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{array} \quad \text{and} \quad \begin{array}{cccc} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{array}$$

let's say $j(K)$ and $j'(j(K))$, respectively.

Hence the requirement that the players' choices should be invariant under these 'inessential' transformations is captured by the following:

Transformation principle
Let $K \in \wp^+(2^A)$, and j be a transformation of K. Then

$$j(R(K)) = R(j(K)). \tag{Tr}$$

Intuitively, the Transformation principle states that applying some transformation j to the set of best elements (according to R) of K is just the same as choosing the R-best elements of the transformation of K by j.

The second step then in the construction of the Minimum Ambiguity Reason consists in the formalization of the 'ambiguity of worlds within K', so that agents, while satisfying the Transformation principle will go about selecting the most outstanding elements of K—the focal points. Notice that, as one would clearly expect from the discussion on triangulation and focal points, 'ambiguity' is being characterized as a contextual notion, relative in fact to the knowledge K.

So let $K \in \wp^+(2^A)$. Then for $f \in K$, *the ambiguity class of f within K at level m* is recursively defined by:

$$\mathbb{S}_0(K, f) = \{g \in K \mid \exists \text{ trans. } j \text{ of } K \text{ such that } j(K) = K \text{ and } j(f) = g\}$$

$$\mathbb{S}_{m+1}(K, f) = \begin{cases} \{g \in K \mid \ |\mathbb{S}_m(K, f)| = |\mathbb{S}_m(K, g)|\} & \text{if } |\mathbb{S}_m(K, f)| \leq m + 1; \\ \mathbb{S}_m(K, f) & \text{otherwise.} \end{cases}$$

The intuition of the base case is that of grouping together those possible worlds g which are in the range of a transformation j of K to itself taking f as argument, thus giving an initial measure of the ambiguity of f in K. The recursive step, on the other hand, causes worlds with the same ambiguity to be grouped in the same class, the purpose of the side condition being that of avoiding coalescing classes 'too quickly' (and hence possibly losing some 'natural' features of the relevant classes).

Define now, for $f, g \in K$, the relation

$$g \sim_m f \Leftrightarrow g \in \mathbb{S}_m(K, f).$$

Recall that one of the requirements of the algorithm is that agents should avoid selecting one but not both elements of a pair of undistinguishable options. Indeed the following proposition ensures that as f ranges over K, $\sim_m$ induces a partition on K.

Proposition 1. *$\sim_m$ is an equivalence relation and the sets $\mathbb{S}_m(K, f)$ are its equivalence classes.*

Moreover, this m-th partition is a refinement of the $m + 1$-st partition. In other words, the sets $\mathbb{S}_m(K, f)$ are increasing and so eventually constant fixed at some set which we shall call $\mathbb{S}(K, f)$.

We are now ready to introduce the *ambiguity of f within K*, which is formally defined by:

$$\mathbb{A}(K, f) =_{def} |\mathbb{S}(K, f)|.$$

Finally, we can define the *Minimum Ambiguity Reason* $R_{\mathbb{A}}(K)$ by letting:

$$R_{\mathbb{A}}(K) = \{f \in K \mid \forall g \in K,\ \mathbb{A}(K, f) \leq \mathbb{A}(K, g)\}. \tag{1}$$

As an immediate consequence of the definition of $R_{\mathbb{A}}$ we have the following result:

Proposition 2. *$R_{\mathbb{A}}(K) = \mathbb{S}(K, f)$, for any $f \in R_{\mathbb{A}}(K)$*

Recall that agents have to select a *unique* option from K, so as argued when introducing the informal procedure, whenever the size of $R_{\mathbb{A}}(K)$ is greater than 1, players will just randomize.

The following results show that the intuition that players of the conformity game should select the 'most distinguished' worlds from a set K while satisfying closure under undistinguishability is indeed captured by the minimum ambiguity reason.

Theorem 3. *$R_{\mathbb{A}}$ satisfies Transformation.*

Theorem 4. *A non-empty $K' \subseteq K$ is closed under transformations of K into itself if and only if there exists a Reason R satisfying Transformation such that $R(K) = K'$.*

The importance of these results is that in the construction of $R_{\mathbb{A}}(K)$ the choices $\mathbb{S}_m(K, f)$ which were eliminated (by coalescing) because of there currently being available an alternative choice of a $\mathbb{S}_m(K, g)$ of the same size are indeed equivalently being eliminated on the grounds that there is a like-minded agent, even one satisfying Transformation, who could pick $\mathbb{S}_m(K, g)$ in place of $\mathbb{S}_m(K, f)$. In other words it is not as if some of these choices are barred because no agent could make them whilst still satisfying Transformation. Once a level m is reached at which there is a unique smallest $S_m(K, f)$ this will be the choice for the informal procedure. It is also easy to see that this set will remain the unique smallest set amongst all the subsequent $S_n(K, g)$, and hence will qualify as $R_{\mathbb{A}}(K)$. In this sense then our formal procedure fulfills the intentions of the informal description given at the beginning of this section.

3.4 Concluding remarks

We conclude by evaluating the extent to which the Minimum Ambiguity Reason contributes towards providing a formalization of the problems arising in the process of triangulation and in the selection of multiple Nash-equilibria in pure coordination games.

$R_{\mathbb{A}}$ and triangulation. The distinct level of abstraction stands out in the comparison of the radical interpretation and the conformity game situations. While with the radical interpretation problem it is attempted to lay down a theory of interpretation *for natural languages*, the choice problem faced by the agents in the conformity game is based on the selection of otherwise meaningless binary strings. In both cases, however, agents should rationally aim at performing *disambiguating* choices and the framework of Rationality-as-Conformity provides agents with an algorithmic procedure to achieve this. It is a matter of future research to investigate the disambiguation of options arising in gradually more and more complicated structures.

Whilst the agents involved in the radical interpretation situation can appeal to actual *observations* of their own reciprocal (non linguistic) behaviour, the players of the conformity game can only *conjecture* about the expected behaviour of their fellows. Again, we see this as a difference of levels of abstraction, yet not of kind, as we concentrate on the 't_0' of the triangulation process, when the transition takes place from agents not sharing any communication devices, to conforming on the use of some. This is being paralleled by the controlled experiments in pure coordination games, as reported, e.g., in Mehta et al. (1994).

$R_{\mathbb{A}}$ and focal points. How far the Minimum Ambiguity Reason goes towards providing a solution to pure coordination games depends, in the first place, on whether the *uniqueness* of the selection is considered a necessary condition on the solution concept or not. Since the early investigations in focal points and salience, uniqueness has been given considerable importance. In some recent, computationally-oriented investigations on the subject, however, other properties of focal points have received attention, with the uniqueness requirement being considerably relaxed (see Kraus et al. (2000) for a comprehensive study). The construction of the Minimum Ambiguity Reason makes explicit the fact that certain coordination problems might be so nebulous that agents cannot rationally go beyond the selection of 'small' sets of options, the minimally ambiguous ones, if the closure under undistinguishability requirement is to be satisfied. The drawback for failing this being, as illustrated above, the possibility of systematically missing coordination.

Acknowledgments

I am greatly indebted to Jeff Paris for the formulation of the Minimum Ambiguity Reason and for many stimulating discussions on the topic. An early version of this paper was presented at *The 2004 Prague International Colloquium on Logic, Games and Philosophy: Foundational Perspectives*. I wish to thank the participants for many helpful remarks.

References

Camerer, C. (2003). *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton University Press, Princeton, NJ.

Davidson, D. (2001). *Subjective, Intersubjective, Objective*. Oxford University Press, Oxford.

Feldman, R. (1998). Principle of charity. In Craig, E., editor, *Routledge Encyclopedia of Philosophy*. Routledge, London.

Glock, H. (2003). *Quine and Davidson on Language, Thought and Reality*. Cambridge University Press, Cambridge.

Harsanyi, J. and Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*. MIT, Cambridge, MA.

Hosni, H. (2005). *Rationality as Conformity.* Doctoral thesis, School of Mathematics, The University of Manchester, Manchester.

Hosni, H. and Paris, J. (2005). Rationality as conformity. *Knowledge Rationality and Action (Synthese)*, 144(2): 249–285.

Janssen, M. (1998). Focal points. In *New Palgrave Dictionary of Economics and the Law*. MacMillan, London.

Kraus, S., Rosenschein, J. S., and Fenster, M. (2000). Exploiting focal points among alternative solutions: Two approaches. *Annals of Mathematics and Artificial Intelligence*, 28(1–4):187–258.

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.

McGinn, C. (1977). Charity, interpretation, belief. *The Journal of Philosophy*, 74(9):521–535.

Mehta, J., Strarmer, C., and Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination. *The Americal Economic Review*, 84(3):658–673.

Nozick, R. (1993). *The Nature of Rationality*. Princeton University Press, Princeton, NJ.

Osborne, M. J. (2004). *An Introduction to Game Theory*. Oxford University Press, Oxford.

Parikh, P. (2000). Communication, meaning and interpretation. *Linguistic and Philosophy*, 23: 185–212.

Paris, J. B. (1994). *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge University Press, Cambridge.

Paris, J. B. and Vencovská, A. (1990). A note on the inevitability of maximum entropy. *International Journal of Approximated Reasoning*, 4:183–224.

Paris, J. B. and Vencovská, A. (2001). Common sense and stochastic independence. In Corfield, D. and Williamson, J., editors, *Foundations of Bayesianism*, pages 203–240. Kluwer, Dordrecht.

Quine, W. V. (1960). *Word and Object*. MIT Press, Cambridge, MA.

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, 105(430):533–550.

van Rooy, R. (2004). Evolution of conventional meaning and conversational principles. *Synthese*, 139(2):331–366.

Wachbroit, R. (1987). Theories of rationality and principles of charity. *British Journal for the Philosophy of Science*, 38:35–47.