Rationality as conformity

Hosni, Hykel

2005

MIMS EPrint: **2005.38**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

RATIONALITY AS CONFORMITY

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2005

Hykel Hosni

School of Mathematics

Contents

Abstract

We address the problem of characterizing the choice processes of two *like-minded* yet *non-communicating* agents who intend to select, from a finite set of options, *the same* possible world. Hence, we call the resulting framework "Rationality-as-conformity"

Within the scope of our formalisation, in which a choice problem is defined on a non-empty subset of maps from a finite set A to the binary set $\{0, 1\}$, we introduce and investigate three distinct logico-mathematical characterisations of Rationality-as-conformity.

Finally, we discuss the applicability of our framework to problems such as pure coordination games and radical interpretation which are traditionally related to "rationality".

The key characterisation results presented throughout Chapters 3–6 appear in Hosni and Paris (2005), whereas parts of Chapters 7–8 have been submitted for publication.

Keywords: Rationality, reasons, coordination, choice functions, radical interpretation, selection of multiple-Nash equilibria, social choice.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Acknowledgements

I am enormously indebted to Jeff Paris to whom I owe many of the ideas discussed here *and* a lot of good humour. To be supervised by Jeff has been a true privilege.

Many thanks also to George Wilmers, whose commitment to the well-being of postgraduate students has resulted - on very many occasions - in invaluable help.

I am extremely grateful to the audiences and the organisers of the conferences and seminars where I have had an occasion to discuss the ideas collected in this thesis. Many of those talks, as well as many email exchanges, have been sources of precious feedback, criticism and encouragement. Specifically, I would like to thank Luc Bovens, Robin Clark, Mauro di Nasso, Tommaso Flaminio, Franz Dietrich, Wiebe van der Hoek, David Makinson, Peter McBurney, Patrick Paroubek, and Guglielmo Tamburrini. Thanks also to three anonymous referees for *Knowledge Rationality and Action*.

Final and *very special* thanks to Serena and Edo for their steadfast support.

Grants

I acknowledge an EPSRC scholarship covering my PhD fees. (Unfortunately, however, it is the current policy of EPSRC to distinguish between British and other European citizens, granting maintenance stipends only to the former). The Italian National Research Council, and in particular the Institute for the Sciences and Technology of Information (ISTI-CNR) in Pisa funded a large part of my PhD work under the form of various fellowships, which I most gratefully acknowledge.

From May to July 2004 my research was funded by the Alexander von Humboldt Foundation, the Federal Ministry of Education and Research and the Program for

the Investment in the Future (ZIP) of the German Government.

Finally I am grateful to the School of Mathematics, the Association for Symbolic Logic, the Italian Institute for Advanced Mathematics (INdAM) and the Agent Link for their conference grants.

Chapter 1

Yet another characterisation of rationality

Abstract: *We introduce, motivate and justify the main ideas and concepts underlying the Rationality-as-conformity framework.*

1.1 What is Rationality-as-conformity all about?

Rationality-as-conformity begins with the idea that a "rational", "commonsensical", "natural", or simply "logical" choice is one which corresponds to the choice other similar agents would come up with in similar situations. Our aim is to model the choice processes leading to this sort of conformity. Consider the following example.

Example 1 *(Supermarket shelf arrangement).* There are numerous ways in which a supermarket manager might choose to arrange the shelves in her store, for example by alphabetical order of product name, by product size or weight, by price, by the package's colours, and so on indefinitely (not to mention the astronomic number of random orderings!). However when stepping into a new supermarket (i.e. one we have never visited before, and about which nothing is known to us, apart from the fact it *is* a supermarket) we *expect* to find teas close to coffees, pastas close to rices, nappies near to toilet rolls. At least we argue that it would surely seem *natural* to hold expectations of this sort. In fact, if after ten minutes searching we finally

located the sugar among the washing powders, we might well be inclined to question the store manager's rationality! After all, we see this as a situation where, for mutual convenience, the store manager and ourselves are trying to conform on the selection of a common world, i.e. shelf arrangement.

Although this *is* the sort of situation we intend to model within our framework, it is not hard to see how rapidly the complications would arise, if we were to work with this informal problem. For instance it could be put forward that in fact there is no choice process to be modelled there, just the appropriate use of common knowledge. The objection here would be that there are in fact rules or conventions (discovered in all probability within marketing research) that regulate what a "rational" shelf arrangement is. [Surely there seems to be some cold-blooded form of logicality when it comes to shelving sweets right at the eyes-height of an invariably bored child queueing at the till!] Hence, the objection would conclude, those rules or conventions are all that a "rational" customer needs to learn to shop conveniently. This would surely work if there were something like a "universal shelving rule" around. Yet, needless to say, this is utterly unreasonable. Therefore it is not hard to see that this possible objection just begs the question for "the new customer" would still have to figure out *which* shelving convention the supermarket manager is in fact adopting.

The next objection then, might be to notice that supermarket managers might indeed fill up the store with signs and maps indicating to the unlearned customer where is what. An account of "rational shelving" pursuing this line, however, would seem to be easily exposed to a very basic shortcoming: What if the customer doesn't happen speak the language(s) chosen by the manager for the sings? What if the manager and the customer didn't in fact share any language at all?

In this thesis we shall consider an idealised and mathematically abstract situation where communication or knowledge of such rules and conventions are not available to the agents. Since it is assumed that agents' ultimate goal is that of conforming to the expectations of their peers, we shall refer to the overall approach as Rationality-as-conformity.

Rationality (or *common sense* – we shall not make a distinction among these terms) is, in its full generality an extremely complex and widely debated subject. Yet within the scope of our simple mathematical formulation we shall be able to provide what amounts to a three-fold characterisation or *definition* of what it means to choose rationally. The hope is that such investigations will ultimately provide a way of viewing and understanding these notions in a much more general real world context.

1.1.1 Summary of the thesis

The thesis is organised as follows. The remainder of this chapter is devoted to tracing the motivations for the study of the Rationality-as-conformity and to illustrating in some detail the main problem and the corresponding assumptions. We then move on in chapter 2 to recall the main mathematical characterisations of rationality that directly relate to Rationality-as-conformity in order to point out their general inadequacy to provide a solution for our main problem. The proper characterisations of Rationality-as-conformity are introduced in full detail throughout chapters 4–6, based on the formalisation of the framework given in chapter 3. In chapter 7 some variations on the main theme are considered, whereas in chapter 8 we discuss the applicability of the Rationality-as-conformity framework to the problems of selecting *focal points* in pure coordination games as well as facilitating *triangulation* in radical interpretation problems. Chapter 9 concludes the thesis.

1.2 Probabilistic Common Sense

Suppose that an agent is required to assign a subjective degree of belief to events about which she only has partial information. As usual we assume that the events are represented by the sentences $\theta, \phi, \ldots \in SL$ built up from some propositional language L in the usual way. Within the subjectivistic (Bayesian) picture, such degrees of belief are operationally quantified in terms of betting quotients: the degree of belief an agent has in the sentence (representing a certain event) θ being interpreted in

terms of the agent's willingness to bet on the truth (occurrence) of θ. The resulting betting scenario – introduced independently by de Finetti (1931) and Ramsey (1964) at the end of the 1920's – leads to a definition of a rational (or coherent, or consistent) assignment of degrees of belief in terms of *fair betting quotients*, that is to say, assignments that prevent the agent from incurring into sure loss (in a certain series of bets). The simple, yet fundamental intuition here is that it would be irrational, incoherent, or simply *illogical* of the agent to bet in such a way that she would loose under any circumstances, that is, no matter what the outcome of the bet itself will be.

The cornerstone of the subjectivistic-Bayesian characterisation of rational belief consists in the so-called *Dutch Book Theorem* (first stated by Ramsey and proved (independently) by de Finetti (1931)), according to which a necessary and sufficient condition for betting quotients to be fair, and hence for degrees of belief to be coherent, is that they satisfy the standard (Kolmogorov) axioms for probability functions. In other words if agents are to avoid sure loss – blatantly irrational behaviour – they must choose degrees of belief according with the laws of probability.

Despite the criticisms related to the underlying assumptions of the betting framework, this powerful theorem lies at the heart of the justification for taking rational degrees of belief as coherent degrees of probability. It is surely not the only argument for "belief as probability", though. Notably Cox's Theorem (Cox, 1946) and related results by Aczel (1966) provide strong support to this view, as discussed at length in (Paris, 1994, ch. 3) where a rigorous reconstruction of Cox's result is given.

But does coherence (so construed) exhaust the intuitive notion of "rationality"?

For de Finetti there was no question about that for everything that probability – the "logic of the uncertain" – can do for the agents' choices is to fix, through coherence, the boundaries of their (ir)rationality. Within those boundaries however, any choice of specific degrees of belief is still permitted. In his fundamental monograph on the theory of probability (de Finetti, 1974, p.109) he notices that, whenever $x_i, i = 1, \ldots, n$ is a set of logically independent events, *any* probability assignment $p_i, i = 1, \ldots, n$ such that $0 \leq p_i \leq 1$, will be coherent and, as far as the betting

framework goes, *rational*. Indeed, de Finetti seemed to believe that any refinement beyond coherence would have resulted in a commitment to some *adhockeries*:

> Whether one solution is more useful than another depends on further analysis which should be done case by case, motivated by issues of substance, and not –as I confess to having the impression– by a preconceived preference for that which yields a unique and elegant answer even when the exact answer should instead be *any value lying between specifiable limits*. (de Finetti (1974) as quoted by Coletti and Scozzafava (2002))

1.2.1 Rational degrees of belief beyond coherence

What de Finetti seems to claim is that beyond coherence there is only *pragmatics*. However, fixing coherence, there are surely cases in which after some reflection, certain distributions of probabilities appear to be more "commonsensical" ("natural", "logical", "obvious" etc.) than others. To see this in a special case, consider the following simple example (Paris, 1994, p.67).

Example 2. Suppose that an agent i knows nothing, so her knowledge can be represented as $K = \emptyset$ and let $L = \{p_1, p_2\}$ be a propositional language. The agent i is asked to give a value to $bel(p_1 \vee p_2)$ under the assumptions that bel is a subjective probability function (on SL) and that K *is* everything i knows (the Watts Assumption of Paris (1994)). It is immediate to see that also in this case, as far as coherence goes, any value between 0 and 1 would do for i. Still there seems to be a way of reasoning that refines such an "uninformative" suggestion about the probability to be assigned. For, if i knows nothing, she has no reason to prefer certain possible valuations on L overs some others, that is to say that the atoms

$$p_1 \wedge p_2, \neg p_1 \wedge p_2, p_1 \wedge \neg p_2, \neg p_1 \wedge \neg p_2$$

of SL should all be assigned probability 1/4, given that there is no grounds for distinguishing them apart and that the sum of their probability values must add up

to 1. Using this bit of "structural" information, together with the fact that

$$\models p_1 \vee p_2 \leftrightarrow (p_1 \wedge p_2) \vee (\neg p_1 \wedge p_2) \vee (p_1 \wedge \neg p_2)$$

the agent will be lead assign her probability accordingly, i.e.:

$$bel(p_1 \vee p_2) = bel(p_1 \wedge p_2) + bel(\neg p_1 \wedge p_2) + bel(p_1 \wedge \neg p_2) = 3/4.$$

The study of commonsensical *inference processes* developed over the past 20 years by Paris and Vencovská addresses the problem of refining, by means of a small number of "common sense principles" (similar to the ones implicitly applied in the example above) the notion of subjective coherent assignment of probabilities subject to the constraints imposed by the knowledge K possessed by an agent.

The emphasis of this characterisation - indeed a distinguishing feature of this approach - is on the fact that "rationality" or "common sense" is being formalized by specifying the desired properties of the reasoning process rather than by specifying the desired features of its outcome. This approach, which could be termed *process-based*, implies that an individual agent fails to be rational if she fails to adhere to (some of) the principles which are identified with common sense. This contrasts with the approach which could be called the *outcome-based* according to which irrationality is synonymous with the selection of a sub-optimal option, where optimality is usually characterised in terms of some utility function. As we shall note in more detail throughout the following chapter the distinction between process- and outcome-based characterisations constitutes an important ground of comparison between alternative characterisations of rational choice.

For the sake of keeping the discussion self-contained, we recall now the main elements of the Paris-Vencovská characterisation, starting with a very informal description of their Common sense Principles:

Renaming Changing the names things are called should not result in agents changing their assignment of probabilities.

Obstinacy Learning information already possessed by an agent should not result in her changing her mind.

Irrelevance Agents should ignore knowledge that is known to be "irrelevant" (where this is formally defined) to the problem at hand.

Equivalence In the presence of two identical knowledge bases should engender the same probability assignments.

Continuity Microscopic changes in the knowledge base possessed by an agent should not cause macroscopic changes in the probabilities assigned.

Relativisation The probabilities that an agent would be willing to assign on the occurrence of a certain event should only depend on the knowledge one agent would have if that event occurred.

(Weak) Independence Conditional beliefs interpreted as conditional probabilities should satisfy a (weak) notion of statistical independence.

In the case of knowledge being represented in terms of linear constraints on a subjective probability function, the fundamental result of this characterisation goes as follows:

Theorem 1.1 (Paris and Vencovská (1990, 1997)). *If the above set of principles is adhered to, then an agent's assignment of probability values on the basis of a knowledge base K is completely determined, for all K. That is, there is only one probability distribution that is consistent with a given K, the one given by the Maximum Entropy inference process.* ■

The remarkable feature of commonsensical probabilistic reasoning that emerges from this framework is that the requirement of adherence to the common sense principles determines a unique way of assigning degrees of belief to the sentences in SL. So, despite being "process-driven", this characterisation has a deep, if indirect, impact on the actual probabilities that agents should assign. As an immediate consequence of this, if distinct agents possess essentially the same knowledge *and satisfy common sense*, they must end up assigning essentially the same degrees of belief to the as yet undetermined sentences of their language. This is in fact a normative requirement

that implies that *agents must conform if they are to be "rational"*. This gives us a first important justification and motivation for taking Rationality-*as*-conformity.

Note that this is in consonance with the subjective-probabilistic tradition. Suppes, to take a remarkable example, suggested that

> the task of the theory of rationality, for the Bayesian, is to understand how to conceive and design experiments that will eliminate or reduce diversity of opinion about serious questions, and part of the task of this theory is being clear about puzzling matters like the paradoxes of confirmation. (Suppes, 1966, p.204)

We conclude by noting, as discussed in full detail in Paris (1999), that the Paris-Vencovská characterisation is based on the simple yet fundamental idea that choice, if rational, must be grounded on (good) reasons. This idea materialises in the fact that commonsensical agents who have no grounds for *distinguishing* among a pair of options should not be willing to prefer one option over the other, for they would fail to have good reasons to do so. Hence, we can see that the guiding idea for the formalization of commonsensical inference processes consists in constraining, via principles, invariance under such an indistinguishability. [Note that this relates to some defences of (a suitably formulated version of) the principle of indifference and notably that of Jaynes' (see, e.g. Jaynes, 1979).]

The main goal of the Rationality-as-conformity framework is to account for this very essential feature of rationality in what is arguably the simplest choice situation consistent with this intuition.

1.3 Introducing Rationality-as-conformity

An upshot of the Paris-Vencovská characterisation is that two agents who(se inference processes) satisfy the principles of common sense recalled above and who share essentially the same knowledge, must end up assigning essentially similar degrees of belief.

Now, in practice, we can look at the assignment of degrees of belief as a problem of choice defined on possible probability distributions. In fact this correspondence between assignments of probabilities and choices is often regarded as central in the formalization of reasoning under uncertainty. A representative example of this position is given by de Finetti, who remarks in his discussion of *proper scoring rules* as ways of eliciting probabilities that

> the choice of a particular action among a sufficiently wide set of permitted possibilities is *equivalent* to an evaluation of the probability concerned. (de Finetti, 1972, p.20, added emphasis)

and he refers to this as a "well-known conclusion of decision theory".

In this spirit then, we can ask what sort of *choice process* should be adopted by two agents who, sharing essentially the same way of reasoning yet being otherwise mutually inaccessible, intended to select the same "world" from a given finite set of possible ones. It is the choice of the same possible world that we shall identify here with *conformity*.

If possible worlds coincided with probability distributions, the probabilistic commonsensical agents of the Paris-Vencovská characterisation would have no choice other than the distribution with the largest possible entropy. What we undertake to investigate with Rationality-as-conformity is the formalization of the choice processes by means of which conformity can be achieved in a framework in which much weaker assumptions are made about the nature of the knowledge possessed by the agents. In fact we will move from the representation of knowledge given in terms of consistent sets of linear constraints on a subjective probability function all the way up to non-empty subsets of all the maps from a set A to a set B.

Before considering some examples of Rationality-as-conformity, we need to make a little bit more precise the interpretation of our main problem as well as the main assumptions on which the entire framework depends.

1.3.1 The Main Problem and the assumptions

As already anticipated, the main problem addressed by the Rationality-as-conformity framework is that of characterizing the choice processes of two *like-minded* yet *non-communicating* agents who intend to select, from a finite set of options, *the same possible world*. In particular, agents are not assumed to have any other goal or intention (and corresponding 'beliefs') than achieving conformity. And as we shall see in full detail, the mutual expectations of conformity constitute the only value for possible worlds.

In our attempt to provide firstly a formalization and then a solution to the main problem, we shall commit to some assumptions. We regard the minimality of those assumptions (in terms of number and in terms of strength) as a distinctive feature of Rationality-as-conformity.

Process-based perspective

A key standpoint of our framework is that "rational choice" is to be characterised in terms of the *process* by means of which a choice is arrived at, rather than in terms of its outcome. We shall call this the *Process-based perspective*. We have just remarked that this approach is distinctive of the Paris-Vencovská characterisation. It is, however, relatively uncommon in the "conventional" mathematical theories of rational choice, like decision or game theory (social choice theory, as we shall illustrate later on, can be consider to be an exception to this). In Nozick's account of "rational belief", on the other hand, this assumption plays a central role:

> The rationality of a belief may derive from the process by which that belief is arrived and maintained, but not every (conceivably) effective way of arriving at true belief would mark a belief as rational. [...] [R]ationality is not simply *any* kind of instrumentality. It requires a certain type of instrument, namely reasons and reasoning. Suppose, then, that a particular procedure is a reliable way to arrive at a true belief. If an action or belief yielded by that procedure is to be rational, not only must the

> procedure involve a network of reasons and reasoning, but this also must be (in part) *why* the procedure is reliable. The reasons and reasoning contribute to the procedure's reliability. (Nozick, 1993, p.71)

Although Nozick's passage concerns specifically the relation between the "rationality" of a belief and its "truth" (investigating the possibility of conceiving a false albeit rational belief), it is entirely consonant with Rationality-as-conformity. It is in fact in this spirit that we formalize rational choice in terms of *reasons*.

The Fundamental assumption

This brings us to what is perhaps the single most important assumption of Rationality-as-conformity. Though widely endorsed in the mathematical characterisations of rationality, it rarely receives explicit mention. It is the assumption that unless certain conditions apply, choosing randomly is not "better" than choosing according to some reason, where the latter is intended as an adequate criterion (adequate, that is, to the achievement of the agent's goals). We refer to this as the *Fundamental assumption*.

The importance of the Fundamental assumption for the study of rationality is, as Simon puts it, "outside dispute":

> Everyone agrees that people have reasons for what they do. They have motivations, and they use reason (well or badly) to respond to these motivations and reach their goals. (Simon, 1986)

Of course, this assumption does not imply that in the characterisation of rationality there can be no space for "random" choices, i.e. choices performed by *picking* one option according to the uniform distribution. There are situations, in fact, in which this is the only advisable strategy. At the most abstract level, those situations will occur whenever an agent faces a set of options which, apart from being distinct, are otherwise completely indistinguishable. In economics, to make a more concrete example, it is generally accepted that in those games with Nash-equilibria in *mixed strategies*, rational players *should* randomize. Finally, in autonomous robotic navigation, many examples are found of situations in which the best way for a robotic agent

to avoid the obstacles that hinder its navigation involves randomizing its trajectory (see e.g. Ram et al., 1997; Arkin, 1998).

If all those examples make a clear point for the respectability of random choice in the characterisation of rationality, it must be appreciated that this must be subject to the satisfaction of certain specific conditions. These could be, in the above examples, the agent's failure to distinguish among options; the fact that a matching pennies player's choice of "heads" should be "as unpredictable as possible" for the opponent; the fact that the obstacles that hamper the robot's navigation are tightly cluttered, and so on. We shall see that Rationality-as-conformity gives its place to randomization as well.

Relativisation of rationality

We have already remarked that Rationality-as-conformity is heavily inspired and motivated by the Paris-Vencovská characterisation. In fact the former can be seen as an attempt to account for the normative requirement imposed by the latter, namely that agents facing certain choice situations – if commonsensical – should conform. Hence, it is immediate to appreciate how our main problem leads to a relativized characterisation of an agents' rationality, relative, that is, to the choices performed by the others. We shall refer to this as the *Relativization of rationality* assumption.

Note that since we will also be assuming that agents cannot communicate, this relativization makes the kind of interaction between the agents genuinely strategic. Indeed, Relativization of rationality plays a primary role in many areas of the social sciences and is surely one of the cornerstones of the theory of games (we shall discuss this more extensively later on).

In epistemology, a counterpart of it can be found again in Nozick's account of rationality. He argues that

> Sometimes it will be rational to accept something because others in our society do. Consider the belief mechanism that brings you to accept that what (you can see) most other people believe. We are all fallible, so the

> consensus of many other fallible people is likely to be more accurate than my own particular view when it concerns a matter to which we all have equal access. For a wide range of situations, the mean of a larger sample of observations is likely to be more accurate than one randomly selected individual observation. (Nozick, 1993, p. 129)

An interesting aspect of this sort of justification for the Relativization of rationality is that it pivots on the fallibility of agents. In other words it is the *bounded rationality* of the individual that justifies her in revising her beliefs in the event of blatant disagreement with the majority's view. An important difference between this idea and the Relativization of rationality, however, must be emphasized, namely the fact that in the latter agents relativize to the *expected* choices of the others rather than to their actual behaviour.

Another interesting analogue to the Relativization of rationality can be found in Keynes discussion on the "investors" and "speculators" in financial markets:

> [P]rofessional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practise the fourth, fifth and higher degrees. (Keynes, 1951, p.156)

Interestingly, however, the kind of relativization underlying the "beauty contest"

is intended by Keynes as directed towards adopting a *minority* behaviour, rather than a majority one. What the "clever" investor aims to do, in fact, is to outperform the majority by, say, selling shares just before (she thinks that) everyone else will start selling, hence maximizing the profit.

Introspective agents

The Relativization of rationality is closely connected with another assumption on the nature of the agents featuring in the Rationality-as-conformity framework, namely the fact that agents are capable of introspecting. This, which we shall refer to as the *Introspective agents assumption*, is surely one of the key abstractions of the entire framework. It amounts to assuming that agents have full access to their own options and that they never make mistakes when doing that. It goes without saying that these idealisations can easily fail in the "real world". Nevertheless, given the nature and the goals of our present analysis, we find this abstraction entirely acceptable.

Common knowledge

The last general assumption of Rationality-as-conformity consists in the fact that the agents have common knowledge of the mathematical structure of the choice problem that they are facing and common knowledge about each other's intention to conform to their mutual choice expectations. Naturally enough, we call this the *Common knowledge assumption*.

Again, we endorse this assumption, as it is usually done in the theory of games, despite the fact that it can exceed the powers of boundedly rational agents.

While the assumptions illustrated so far can be viewed as a set of maxims, epistemological or methodological, that underlie the general characterisation of Rationality-as-conformity, the set to be introduced below captures the *specific assumptions* that we make in the remainder of this work. Specific, that is, to the choice situation of our main problem. The intuition behind distinguishing between these two sets of assumptions is, fixing the general ones, that of being able to modify the latter in order to apply the Rationality-as-conformity framework to various sorts of choice problems.

As it will become clear later on, our three specific assumptions provide the main guidelines for the formalization of Rationality-as-conformity to follow.

Inaccessibility

Given that our main problem aims at capturing the most basic (that is to say simplest, least structured, etc.) situation consistent with the intuitions of Rationality-as-conformity, we shall start by assuming that agents don't know anything specific (that is beyond what is entailed by Common knowledge) about each other's way of structuring the world. Moreover, we shall also assume that communication among them is not allowed. Hence, putting these two together, we shall commit to what we call the *Inaccessibility assumption*.

Inaccessibility ties the main problem to a number of situations which are widely studied, from political science to the theory of strategic non-cooperative games, to the theory of multi-agent systems. In each of these areas motivations can be found for assuming that "rational" agents might have to operate in the absence of communication, and still must be able to conform (in our terminology). Indeed, in many situations it is advantageous for agents to refrain from communicating, as the exchange of relevant information might be unreliable, unsafe, or simply too expensive (compared to the resulting benefits).

Communication-less scenarios are of fundamental importance in the area of political science concerned with the so-called *strategy of deterrence*, of which an early and very influential account was given by Schelling (1960). The problem of the unreliability of communication in coordination problems is studied extensively in the distributed and multi-agent systems literature (see, e.g. the paradigmatic example of *coordinated attack problem* Halpern et al. (1995)). It is folklore in economics, on the other hand, that the open exchange of information can be risky and hence should be avoided in many strategic situations, say when firms operating in oligopolies have to decide their price policies. Finally, agents might have in principle the possibility of exchanging information in order to facilitate conformity, yet in practice this would just be too onerous. Interesting examples of this situation, mainly from the distributed-

and multi-agent systems literature, are described at length in Kraus et al. (2000). Situations of this sort, which arise commonly in the so-called *coordination problems*, all call for a formal study of rational choice behaviour in the absence of accessibility among agents.

Note that Inaccessibility implies that agents cannot agree to adhere to (otherwise arbitrary) *conventions* for the simple reason that they do not posses a shared language in which to stipulate agreements of this sort. Hence, Rationality-as-conformity can be utilized as a framework to investigate the *origin* of (spontaneous) convention.

Likemindedness

Yet our aim here is that of modelling the choice processes that lead agents to conform on the selection of a possible world. Inaccessibility implies that there is no "specific knowledge" that agents have about each other, like, for instance, their past behaviour in similar situations, or information about their general preferences. All the actual information they have is captured by the Common knowledge assumption. Yet, in order to be able to wedge into each others' minds, the agents involved in the conformity problem must have some "structural" information about their peers. To this effect, what seems to be the weakest assumption consists in informing the agents that they are facing other similar agents, where similarity roughly refers to the way the agents "see the world" and "reason about" it. We shall refer to this as the *Likemindedness assumption*.

Of course it is by no means easy to specify what it really means in the "real world" to share the same way of reasoning. Therefore we will be in a better position to appreciate this point once a mathematical formulation of Reasons (choice processes) will be available. However, we have already seen this concept in action in the Paris-Vencovská characterisation, where the agents' reasoning is captured by the notion of a commonsensical inference process. In such a framework Likemindedness mainly concerns the fact that agents who share the same views on what it is commonsense must end up assigning essentially similar degrees of belief. This result, clearly, gives us an initial motivation to consider Likemindedness.

There is, however, another reason – which could be termed operational – to endorse this assumption. Suppose that two agents are involved in a conformity problem. If they don't know anything about each other's way of reasoning, it can be argued that unless they collect strong evidence to the contrary, they have no reason to assume that there are fundamental differences in their "world-views". An argument of this sort, captured by the so called Principle of Charity, lies at the heart of the discussions on *radical interpretation*. This latter, very intuitively, amounts to accounting for the process that leads two individuals, each with their own view world yet who do not possess a shared language, to establish communication. The deep connections between Rationality-as-conformity and radical interpretation will be discussed later on in section 8.2.

Saliency

Our last assumption can appear to be a more or less direct consequence of the previous ones, and is the assumption that agents will indeed select a given option x on recognition of the fact that x appears to be an outstanding element within the set of possible worlds which defines the current choice problem faced by the agents. We refer to this as the *Saliency assumption*.

The reason for endorsing Saliency is as follows: given Introspection, agents will realize if among the set of possible worlds under consideration there exists some option which stands out in comparison to the others. Likemindedness and Common knowledge, on the other hand, will support the expectation that if one such option is recognised as outstanding by an agent i, so it will be for his fellow j. And j will expect that i expects this, and so on. At this point, given Inaccessibility, the outstanding option looks to both agents as the "obvious" choice to be made in order to facilitate conformity.

In the light of Saliency, we can see that as far as Rationality-as-conformity goes, the "rational" choice amounts to what we might informally refer to as the "natural", "obvious", or even "logical" choice to make. Although Saliency can be justified in the grounds of the previous principles, we consider it as an independent assumption.

A discussion of closely related concepts will follow in section 8.4.

1.4 Motivating Rationality-as-conformity

We have insisted that Rationality-as-conformity is indeed inspired by everyday considerations of rationality coinciding with the informal use of the expression "rational", "commonsensical", "intelligent" or "logical" choice. In this section we list some more examples of Rationality-as-conformity emphasising its various aspects. The fact that the structure of Rationality-as-conformity captures a wide class of interesting examples is clearly a further motivation for the investigation of the corresponding problem. However, as pointed out at the end of this section, we will be able to propose that a new formal framework is required in order to provide a general characterisation of Rationality-as-conformity.

1.4.1 Some examples of Rationality-as-conformity

Example 3 *(Robotic Rendez-vous).* 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 at some location `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 so as to facilitate their meeting? That is, how should they *choose* `l`?

Example 4 *(Keywords selection).* It is common practice in the production of scientific literature to add a small set of keywords to the papers submitted for publication. The problem of selecting which keywords are appropriate (for a given paper) is clearly a problem of achieving conformity. It seems, in fact, that a rational (natural, obvious, logical, etc.) way for an author to get round this problem is to introspect and guess which keywords a potential reader would type-in, say in a database search engine, if he intended to retrieve exactly the kind of paper the author is submitting. Note the

complete symmetry of the situation. It is likewise in the best interests of the reader himself to conform to the author's choice of keywords. Indeed, a natural strategy for him to adopt in the selection of the keywords for his search, is to guess which keywords would he choose, were he to be the author of the sort of paper he is looking for.

Example 5 *(Establishing communication).* Embarking into the business of communicating (verbally) with the others is essentially a problem of Rationality-as-conformity. Take, as in the famous mental experiment of *radical interpretation* two individually "rational" agents who share no language whatsoever. For definiteness we can think of one agent as the interpreter and the other as the interpretee. Suppose further that both agents are willing to establish communication, that is to say that the interpretee is willing to be understood by the latter, whereas the interpreter is willing to understand the former. All this is assumed to be common knowledge, yet nothing else and specifically neither the "mental states" nor the "linguistic habits" of the two agents can be assumed to be common knowledge.

In this form, the problem of conformity is, for each agent, to choose among the possible interpretations of the linguistic utterances those which facilitate mutual understanding. We shall see later on in section 8.2 how closely connected is this problem, and perhaps more generally the problem of language acquisition, to the structure of Rationality-as-conformity.

Example 6 *(Smart usernames).* Rationality-as-conformity can also be a private exercise. Consider the problem of choosing a certain username or password, say for a web service. Agreeing with ourselves in those cases is surely a very logical thing to do! In other words, it seems to be advantageous to select those usernames that we could easily recover by means of introspection: "if I were a logical sort of person, *this* is the username that I would choose"! Notice that this example brings clearly to the foreground the fact that the formalization of Rationality-as-conformity leaves common knowledge out of consideration.

By performing suitable variations on the theme we can see how a broad class

of *coordination*, *classification* and *categorization* problems relate to Rationality-as-conformity. Hence this latter can be brought to bear to a wide spectrum of domains, from economics, to artificial intelligence, to the cognitive sciences.

1.4.2 The need for a new formalisation

Having laid down the main intuitions underlying the problem we wish to investigate it seems natural to ask whether the copious literature on the mathematical formalizations of rational choice doesn't contain already what we set out to find here.

In fact, much of the literature on rational choice is permeated with intuitions that relate to our main problem. Yet it turns out that, to the best of our knowledge, no attempt has been made to provide a unitary characterisation of rational choice behaviour taking into account all those features at once. In a sense, what we seek here is a formal *definition* of Rationality-as-conformity which, in the limited scope of the present framework, would serve as a yardstick for other, more specific, characterisations of rationality.

The purpose of the next chapter is to point out more precisely, if still informally, the nature of those features and locate them among some major existing accounts of rationality ranging from individual choice to strategic choice to social choice. Clearly, there is no claim of completeness in the topics surveyed which in fact have been chosen according to their relevance to our main problem.

1.5 Further comments

The Dutch Book Theorem was stated independently by Ramsey and de Finetti, though the first proof is due to the latter (de Finetti, 1931). Subsequent refinements of the notion of fair betting quotient were discussed by Kemeny (1955) and Shimony (1955). See Paris (1994) for a general proof involving the notions of "strict" fairness and Paris (2001) for a generalization of the theorem that encompasses a variety of possible worlds semantics. Note that "coherence" is used mainly after de Finetti, whereas "consistency" is after Ramsey (1964), "fairness" and "strict fairness" after

Shimony (1955); Kemeny (1955).

Hintikka defines a notion of (ir)rationality which bears a close resemblance to the one underlying the Dutch Book argument:

> What is irrational is the behaviour of a man who would persist in subscribing to an indefensible statement after its indefensibility has been made know to him (Hintikka, 1962, p.109)

The characterisation of the Maximum Entropy inference process as the unique choice of a probability distribution consistent with an agent's knowledge and with the common sense principles outlined above was first given in Paris and Vencovská (1990), and is fully developed in chapter 7 of Paris (1994), culminating in Theorem 7.9. See also Paris (1999) for the unification of the common sense principles under the "Symmetry Principle", and Paris and Vencovská (2001) for the study of the non linear case. Some criticisms to the Maximum Entropy inference process are discussed in Paris and Vencovská (1997).

Chapter 2

A brief excursion into economic rationality

Abstract: *We recall some fundamental accounts of "rational choice" and propose that, despite the many similarities, an adequate solution to the Rationality-as-conformity problem requires a novel framework.*

2.1 A dimension for comparison

There is little doubt that providing an adequate, general and precise definition of what we mean by "rational choice behaviour" in the real world is a daunting task. This seems to contrast, however, with our daily experience: we seem to be more than ready to attach to our peers labels of irrationality, lack of commonsense, illogicality, and the like. In other words, despite the difficulties in providing a neat definition of rationality, we don't seem to be too bad at spotting the lack of it. Hence, we could think of replacing the question of defining "what is rational behaviour" with the more operational "how to assess the rationality of an agent's behaviour".

Once we rephrase the issue in these terms, we seem to have two alternative ways of answering: we can either judge an agent's rationality in terms of the *outcome* of the choices she makes (or is willing to make), or we can evaluate the *process* that she has adopted (or is willing to adopt) when performing her choices. Naturally enough, we

can refer to accounts of rationality that mainly focus on the former as *outcome-based*, while attaching the denomination *process-based* to the latter.

Indeed, we have already made use, if implicitly, of this distinction in the above discussion of the Paris-Vencovská characterisation and we have explicitly mentioned Process-based as a general assumption of the Rationality-as-conformity framework. The reason for formulating it explicitly at this point is that it will help us emphasising a major point of departure of the Rationality-as-conformity framework with respect to the conventional accounts of rational choice to be discussed in the next sections.

The concept of "rationality" is one of the most fundamental in economics, and economics-related research. As Sugden puts it:

> In mainstream economics, explanations are regarded as 'economic' to the extent that they explain the relevant phenomena in terms of the rational choice of individual economic agents. (Sugden, 1991)

In a somehow critical vein, Hammond (1997) admits that "rationality is one of the most over-used words in economics", while Rubinstein finds the very word rationality "mystical and vague" (Rubinstein, 1991, p.923). As noted by Simon (1986) however, "Economics has almost uniformly treated human behavior as rational".

But what exactly counts as 'rational (choice) behaviour' economics? In line with the previous remarks, we can say that an almost universal feature of the economics-related approach is the outcome-based characterisation of rationality. In other words, an agent's rationality is assessed in terms of the properties of the outcome produced by her choice or decision (we shall treat the latter terms as synonyms throughout this thesis). As a consequence, different desiderata on the outcome of the agents' choice give rise to distinct outcome-based accounts of rational choice. Yet there seems to be little pluralism in the choice of the desiderata. There is in fact a neat convergence on the requirement that the outcomes of a rational agent's choices should lead to what is traditionally recognised as *the pursuit of the maximisation of the individual's self-interest*. And given that this latter is the target of a somehow broad class of formalizations of rational choice behaviour, we shall generally refer to this approach

as *economic rationality*.

Without delving into any of the subtleties of the economic account of rationality, we shall outline in this chapter the main aspects of this conception, stressing those which most directly relate to the idea of Rationality-as-conformity.

2.2 The conventional theory of rational choice

We begin our excursus with a fundamental outcome-based characterisation of rationality. Its distinctive trait is the idea that in order to maximise the pursuit of her own individual interests, an agent should have preferences over possible courses of actions which satisfy certain consistency principles. A first exhaustive development of this interpretation is to be found in Savage (1954) who combines the subjective interpretation of probability pioneered by de Finetti and Ramsey, with the axiomatization of utility developed in von Neumann and Morgenstern (1944).

Savage's work is so fundamental and influential for economic rationality that economists have been referring to him as the "best spokesman for conventional rational-choice theory" (Sugden, 1991). In what follows we shall conform to this view by taking Savage's account as the representative of the 'conventional, theory of rational choice'. In fact, as we shall briefly point out in section 2.2.1, its centrality goes well beyond the traditional domain of decision theory.

To frame our discussion, let's start by recalling the fundamental elements of the conventional theory. Usually these are spelled out in terms of a *decision situation*, that is to say a tuple $\langle \mathbb{S}, \mathbb{E}, \mathbb{A}, \mathbb{F}, \leq \rangle$, where:

- $\mathbb{S}$ is a non empty set of *states of the world* $s_1, s_2, \ldots$ (assumed to be mutually exclusive);
- $\mathbb{E}$ is the set of *events* $E_1, E_2, \ldots$ (non empty subsets of $\mathbb{S}$);
- $\mathbb{F}$ is the set of *consequences* $f, g, h, \ldots$;

- $\mathbb{A}$ is the set of *acts* $\alpha_1, \alpha_2, \alpha_3, \ldots$ mapping states to consequences, that is

$$\mathbb{A} = \{\alpha \mid \alpha : \mathbb{S} \longrightarrow \mathbb{F}\};$$

- A *preference relation* $\leq$ that individuals have over acts (interpreted as "it is not preferred or indifferent to").

Savage insists that "a consequence is anything that might happen to a person" (Savage, 1954, p.13), so nothing specific needs to be assumed about the nature of the elements of the set $\mathbb{F}$ apart from the intuitive fact that certain consequences may be more attractive than others to an agent, formalized in the definition of "preference among consequences" introduced below. Such a preference constitutes the basis for distinguishing among options in a decision problem. Note in fact that any two acts $\alpha_i, \alpha_j \in \mathbb{A}$ such that $\alpha_i(s) = \alpha_j(s)\, \forall s \in \mathbb{S}$ are taken to be indistinguishable (Savage, 1954, p.14). Hence, fixing the state, acts can be identified with their consequences. This is a fundamental aspect of the whole outcome-based approach: given that preferences are defined over acts, all that agents can take into account in order to be able to distinguish among the possible options are their (expected) *consequences*. As an immediate formal consequence of this, the states of the world are not mentioned in the formulation of the Weak ordering postulate (see below).

Against this background the axiomatization of rational choice takes place by means of "logic-like criteri[a] of consistency in decision situations" (Savage, 1954, p.19). Note that the justification offered by Savage for the acceptance of the following postulates is the "irrationality" that an agent would face as a consequence of failing them (Savage, 1954, p.7). Following Savage's exposition, we shall introduce the required definitions along with the postulates. Notice that any critical assessment of Savage's postulates is beyond the scope of the present work. As usual we shall write $x < y$ whenever $x \leq y$ but not $y \leq x$.

Postulate 1 (Weak ordering) The preference relation $\leq$ is a weak-ordering (i.e. total and transitive).

Postulate 2 (Sure-thing, part I) For all $\alpha_1, \alpha_2, \alpha_1', \alpha_2' \in \mathbb{A}$ and $E \in \mathbb{E}$ such that:

$$\alpha_1 \restriction E = \alpha_1' \restriction E; \tag{2.1}$$

$$\alpha_2 \restriction E = \alpha_2' \restriction E; \tag{2.2}$$

$$\alpha_1 \restriction \mathbb{S} - E = \alpha_2 \restriction \mathbb{S} - E; \tag{2.3}$$

$$\alpha_1' \restriction \mathbb{S} - E = \alpha_2' \restriction \mathbb{S} - E \tag{2.4}$$

and

$$\alpha_1 \leq \alpha_2, \tag{2.5}$$

then

$$\alpha_1' \leq \alpha_2'.$$

Definition (Conditional preference) The conditions (2.1)to (2.4) (relative to a specific event E) are summarized by saying that α_1 *is not preferred over* α_2 *given* E, written $\alpha_1 \leq_E \alpha_2$.

Definition (Null event) An event E is defined to be *null* if $\forall \alpha_1, \alpha_2, \in \mathbb{A}$, $\alpha_1 \leq_E \alpha_2$.

Definition (Preferences among consequences) Let $\alpha \in \mathbb{A}, f \in \mathbb{F}$ and write $\alpha \equiv f$ if and only if $\alpha(s) = f, \forall s \in \mathbb{S}$. Then, $\forall f_1, f_2 \in \mathbb{F}$:

$$f_1 \leq f_2 \iff [\text{if } \alpha_1 \equiv f_1 \text{ and } \alpha_2 \equiv f_2 \text{ then } \alpha_1 \leq \alpha_2].$$

Postulate 3 (Sure-thing, part II) If E is a non-null event with $\alpha \equiv f$ and $\alpha' \equiv f'$, then

$$\alpha \leq_E \alpha' \iff f \leq f'.$$

Definition (Qualitative personal probability) The event E is said to be *not more probable* than E', written $E \leq E'$, if whenever:

1. $f, f' \in \mathbb{F}$ are such that $f' \leq f$;
2. $\alpha(s) = f$ for $s \in E$ and $\alpha(s) = f'$ for $s \notin E$;
3. $\alpha'(s) = f$ for $s \in E'$ and $\alpha'(s) = f'$ for $s \notin E'$

then $\alpha \leq \alpha'$. (Notice that "probable" does not refer here to the usual, quantitative, notion of a probability function or measure.)

Postulate 4 For all $E_1, E_2 \in \mathbb{E}$, either $E_1 \leq E_2$ or $E_2 \leq E_1$.

Postulate 5 (Non-triviality) There is at least one pair of consequences among which the agent is not indifferent, that is $f_1 < f_2$, for some $\langle f_1, f_2 \rangle \subseteq \mathbb{F} \times \mathbb{F}$.

Postulate 6 (Archimedean axiom) If $\alpha_1 < \alpha_2$ and $f \in \mathbb{F}$, there exists a (finite) partition of $\mathbb{S}$ such that, if α_1' agrees with α_1 and α_2' agrees with α_2 except on an arbitrary element of the partition, say x with $\alpha_1'(y) = \alpha_2'(y) = f$ for $x \in y$, then either $\alpha_1' < \alpha_2$ or $\alpha_1 < \alpha_2'$.

Definition $\alpha_1 \leq_E f_1 (f_1 \leq_E \alpha_1) \iff \alpha_1 \leq_E \alpha_2 (\alpha_2 \leq_E \alpha_1)$ whenever $\alpha_2(s) = f$ for all $s \in \mathbb{S}$.

Postulate 7 If $\alpha_1 \leq_E \alpha_2(s)$ $(\alpha_2(s) \leq_E \alpha_1)$ for every $s \in E$, then $\alpha_1 \leq_E \alpha_2$ $(\alpha_2 \leq_E \alpha_1)$.

Savage is then able to prove that whenever an individual's preferences satisfy the above *consistency* postulates, those determine uniquely a subjective probability function and an equivalence class of utility functions by means of which the agent's preference can be represented.

Theorem 2.1 (Savage (1954)). *Postulates 1-7 are sufficient to ensure the existence of a unique real-valued probability function w such that*

$$E \textit{ is not more probable than } E' \iff w(E) \leq w(E').$$

■

Theorem 2.2 (Savage (1954)). *Postulates 1-7 are sufficient to ensure the existence of a real-valued function u defined over the set of consequences $\mathbb{F}$ such that if:*

(i) $E_i, i = 1, \ldots, n$ *is a partition of* $\mathbb{S}$ *and* α *is an act with consequence* f_i *on* E_i *and*

(ii) $E'_i, i = 1, \ldots, m$ *is another partition of* $\mathbb{S}$ *and* α' *is an act with consequence* f'_i *on* E'_i,

then, $\alpha \leq \alpha'$ *if and only if*

$$\sum_{i=1}^{n} u(f_i)w(E_i) \leq \sum_{i=1}^{m} u(f'_i)w(E'_i).$$

Furthermore the utility function u is unique up to a positive linear transformation.

■

Thus, the consistency postulates $1 - 7$ ensure the existence of a utility and a probability function that lead to a definition of an agent's *expected utility* over the set of possible acts which reflects the individual's preferences. This is the mathematical pivot around which the entire outcome-based characterisation of rationality revolves. So, according to this approach, *the rational decision maker who faces potential uncertainty about the consequences of her actions should be choosing as if she were maximising her expected utility.*

Notice that a distinctive feature of Savage's theory is the following chain of dependency relations between probability, preference and choice. Taking the subjectivistic point of view, his conception on probability depends, ultimately, on preference, as implied by the Dutch Book Argument. Preference, in turn, is interpreted in terms of choice. Under the assumption that f and g are the only options, in fact, an agent prefers f over g if, whenever facing the choice, she will select f (Savage, 1954, p.17). Hence, by axiomatizing consistent preferences, Savages formalizes a theory of rational choice (or decision). This fact makes it comparable to the Rationality-as-conformity framework, a comparison which shows the fundamental point of departure of the latter with respect to the "conventional" theory.

Whilst, in fact, the goal of Rationality-as-conformity is the characterisation of the *choice processes* that agents should adopt upon reflection on the choices that they expect the others to expect (and so on) from them, Savage stresses very emphatically that introspective considerations should not intervene in the characterisation of consistent preference:

> I think it of great importance that preference, and indifference, between f and g be determined, at least in principle, by decision between acts and not by response to introspective questions. (Savage, 1954, p.17)

But, as we have just remarked, "decisions between acts" are entirely dependent on the correlated expected utilities. In other words, all that matters for the rationality of an action (under uncertainty) is its (expected) outcome.

The upshot of this, is that the conventional theory of rational choice points to the ordinal comparison of expected utility as the only criterion for *distinguishing* among possible options. Hence, if the agent fails to distinguish options on the grounds of some expected utility, then she fails to have any good *reason* to choose, and hence prefer, one over the other.

Although its fundamental importance can hardly be questioned, the conventional model is the object of countless criticisms, none of which will anyway be discussed here. Rather, we shall insist on the fact that the outcome-based approach to the theory of individual rational choice, is entirely shared by the theory of interactive rational choice, namely the theory of games. This fact is responsible, among others, for the impasse faced, for instance, by the traditional solution concepts for strategic games in the presence of multiple Nash-equilibria, as in the case of (pure) coordination games briefly introduced below in section 2.3. Indeed, as we shall illustrate in section 8.4, the radical change of perspective – which could be considered as an "introspective turn" – operated by the Rationality-of-conformity framework has among its consequences that of contributing towards defining an analytic solution concept for pure coordination games.

2.2.1 Generalised Expected Utility

Recent results by Halpern and Chu add emphasis to the centrality of the outcome-based model of conventional rational choice. Indeed they are able to prove that Savage's characterisation results can be generalised to encompass a variety of outcome-based decision rules other than expected utility maximisation, such as *maximin* and *minimax regret* (Halpern, 2003).

The upshot of their results is a generalization of Theorems 2.1 and 2.2 to the effect that a free choice of a preference relation over acts is allowed. More precisely, given any preference relation over acts – specifically a relation that need not be transitive – this yields a utility function and a plausibility measure representing the *generalized expected utility* rule, where a *plausibility measure* maps events to an arbitrary partially ordered set (Chu and Halpern, 2003, Theorem 3.1). Given that (subjective) probability functions are a special case of plausibility measures, this result shows that the outcome-based characterisation of rational choice is not necessarily a consequence of the probabilistic representation of uncertainty.

Note in the general result of Chu and Halpern the uniqueness (up to positive linear transformation) of the utility function is lost together with, clearly, the uniqueness of the probability function.

2.3 Game theory

The conventional theory of rational choice focuses on individual – non interacting – agents. Yet "rational" agents of the sort economists, social scientists, artificial intelligence practitioners and ordinary people usually refer to, *do* live in highly interactive contexts. In particular it often happens that the rationality of one individual's choices depends essentially on the choices that other (rational) agents simultaneously make. Think, for instance, of the price policies that a firm might adopt in a highly competitive field (say low-cost airlines) or the behaviour of motorists approaching a junction where the traffic light is temporarily out of order.

Modelling social interactions of this sort constitutes one of the goals of the theory

of games. On the other hand we already pointed out in section 1.3 that among the general assumptions of the Rationality-as-conformity framework is the Relativization of rationality, an assumption which is clearly central in game theory. Yet, the conventional theory of games relevant to the problem of Rationality-as-conformity, falls short of providing an adequate framework for the latter. The goal of this section is to illustrate this by outlining the main points of departures of Rationality-as-conformity with respect to the conventional solution concept for one-shot, two-person, non-cooperative games.

Given a strategic choice situation, that is one in which the desirability of a certain action depends, for each individual, on the simultaneous choices made by other inaccessible agents, we can look at the theory of games as the study of the consequences of the assumptions that:

(i) rational agents are utility maximizers;

(ii) it is common knowledge that they are so.

Note that (i) is a very special case of the Likemindedness assumption: the agents are assumed to be similar to the extent that they are assumed to pursue their individual self-interest. In terms of the mathematical characterisation of utility given by von Neumann and Morgenstern (1944), and the subsequent extensions and generalizations, agents are assumed to pursue the maximisation of their (expected) utility. (We can appreciate at this point how Likemindedness constitutes a key aspect of the generality of the Rationality-as-conformity approach.)

Hence, like Savage's account, the characterisations of rational choice behaviour underlying game theory are essentially outcome-based. Osborne and Rubinstein, for instance, put it like this:

> The models we study assume that each decision-maker is "rational" in the sense that he is aware of his alternatives, forms expectations about any unknowns, has clear preferences, and chooses his action deliberately after some process of optimization. (Osborne and Rubinstein, 1994, p.4)

where the "process of optimization" usually is, as already remarked, the maximisation of (expected) utility. In fact, Osborne and Rubinstein (1994) go on by stating that in the *absence* of uncertainty, a complete model of rationality is given by a tuple $\langle A, C, c, \geq \rangle$:

- actions A
- consequences C
- a consequence function c associating an action with a consequence
- a preference relation $\geq$ on the set of consequences (represented via utilities, if needed).

Hence a rational decision maker confronted with a set of feasible actions $K \subseteq A$ will choose an action $\mathbf{a} \in K$ such that $c(\mathbf{a}) \geq c(a)$, $\forall a \in K$.

Before going into some details of the actual characterisation of rationality endorsed by the theory of games, we must restrict somehow the scope of our discussion. In fact, the theory of games is so extremely rich in variety and applicability that it would be extremely hard - if not impossible (Luce and Raiffa, 1957, p.104) - to subsume all the corresponding notions of rationality under a single heading. Thus, for present purposes, we shall confine ourselves to the discussion of a sort of game that relates directly to our main problem, one-shot, non-cooperative, non-zero-sum games.

A *non-cooperative* game is a kind of strategic interaction in which, n players face the problem of simultaneously selecting a strategy from a set of possible ones without being able to stipulate binding agreements (i.e. coalitions) with their opponents. Games of this sort are also referred to as *normal-form* or *strategic*. A two-persons non-cooperative game is *non-zero-sum* if a player's win doesn't imply her opponent's loss (and *vice versa*). In what follows, we shall mainly be referring to this type of game.

Obviously, the mere maximisation of utility cannot constitute a "rational" solution concept for the players of strategic games, as the actual outcomes depend on the

choices made by other, inaccessible, agents. It follows that a natural way of characterizing "rational choice behaviour" for non-cooperative games consists in requiring that each player should select the strategy which happens to be the "best response" to whichever will be the one chosen by his opponent. A pair of such strategies is referred to as a *Nash-equilibrium*, perhaps the single most important concept underlying the conventional theory of non-cooperative games.

More precisely a *strategic game* is defined by

- a finite set N (the set of *players*)
- for each player $i \in N$ a non-empty set A_i (the set of *actions* available to player i)
- for each player $i \in N$ a preference relation $\succsim_i$ on $A = \times_{j \in N} A_j$ (the *preference relation* of player i)

so that a *Nash-equilibrium of a strategic game* $\langle N, (A_I), (\succsim_i) \rangle$, can be defined as a profile $a^* \in A$ of actions such that for every $i \in N$:

$$(a^*_{-i}, a^*_i) \succsim_i (a^*_{-i}, a_i), \quad \forall a_i \in A_i,$$

where a^*_{-i} is the complement of i in N.

If Theorems 2.1 and 2.2 provide the mathematical backbone for the conventional theory of rational choice, the mathematical pivot of the theory of non-cooperative games is provided by Nash's Theorem which guarantees the *existence*, under certain conditions, of at least one equilibrium pair (Nash, 1951).

Hence the theory of Nash-equilibrium accounts for a relativised notion of an agent's rational choices, a feature of Rationality-as-conformity that was clearly missing from the conventional model of rational choice outlined above. Yet it has in common with this latter the fact that – being purely outcome-based – it fails to account for "rational choice" in those cases in which options (i.e. strategies, actions etc.) cannot be distinguished on the grounds of their (expected) utility. In fact, as noted, again, by Osborne and Rubinstein, the solution concept based on the notion of a Nash-equilibrium

> [...] captures a *steady state* of the play of a strategic game in which each player holds the correct expectation about the other players' behavior and acts rationally. *It does not attempt to examine the process by which a steady state is reached.* (Osborne and Rubinstein, 1994, p.14, latter emphasis added)

This limitation, which is perceived by one of the pioneers of the investigations on "bounded rationality" as typical of economic rationality (see, e.g Simon, 1986), can be illustrated effectively by means of a particular class of strategic games, the so-called *pure coordination games*. These make a clear point for a *process-based* characterisation of rational, strategic choice. The main intuition being that whenever outcomes are (utility-)indistinguishable, in fact, only the *process* utilised by the agent to arrive at a choice can account for the rationality of the choice itself.

The next section briefly illustrates the main features of pure coordination games (and consequently the problem that they pose to the traditional theory of Nash-equilibrium). This will allow us to point to a second motivation for taking Rationality-*as*-conformity. Besides the normative precept of the Paris-Vencovská characterisation according to which commonsensical agents *must* conform (see Theorem 1.1), we will find out that conformity (in the case of coordination games) is something that people *can* achieve. In other words, the normative motivation for conformity is coupled with evidence of the fact that the model prescribes that rational agents should follow a pattern of choice behaviour that, in some way or another, they appear to be able to follow "naturally". As Schelling puts it:

> People *can* often concert their intentions or expectations with others if each knows that the other is trying to do the same. (Schelling, 1960, p.57)

2.3.1 Rational choice with multiple Nash-equilibria

Roughly speaking, a *coordination game* – introduced in the game theoretical literature by Schelling (1960) – 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.

One of the classical examples introduced in Schelling (1960) concerns a married couple who get accidentally separated in a supermarket and want to rejoin:

> When a man loses his wife in a department store without any prior understanding on where to meet if they get separated, the chances are good that they will find each other. It is likely that each will think of some obvious place to meet, so obvious that each will be sure that the other is sure that it is "obvious" to both of them. One does not simply predict where the other will go, since the other will go where he predicts the first to go, which is wherever the first predicts the second to predict the first to go, and so on ad infinitum. Not "What would I do if I were she?", but "What would I do if I were she wondering what she would do if she were I wondering what I would do if I were she ...?". (Schelling, 1960, p.54)

Schelling calls this a problem of "tacit coordination" with "common interests". Note that this is a clear example of a conformity problem, that is, essentially similar to those illustrated above in section 1.4.1

A fundamental feature of *pure* coordination games consists in the fact that they are symmetric with respect to payoffs and players. That is, for each individual, any choice among the possible strategies (the supermarket locations in the original example) is "as rational as any other", provided that it conforms to the choice of the other agent. In fact, each point along the diagonal of identical pairs of feasible strategies is a Nash-equilibrium. Thus, given that there are as many utility-indistinguishable equilibria as there are feasible strategies, we must conclude that applying the conventional solution concept for non-cooperative games to coordination problems amounts to no progress whatsoever towards the characterisation of rational choice behaviour: in practice players will be choosing according to the uniform distribution.

Very roughly then, if one assumes that the theory of Nash-equilibrium does characterise rational choice behaviour in strategic games, one must conclude that pure

coordination games admit of no "rational" solution. [We shall put this more precisely later on, when the problem of defining rational choice in pure coordination games will be re-examined in the light of the formalization of Rationality-as-conformity.]

In spite of this discouraging conclusion, as Schelling himself noted through a number of "unscientific" experiments, people do seem to be able to conform on coordination problems with a remarkable rate of success. Extensive empirical investigations over the past two decades, both in the form of controlled experiments (see Mehta et al., 1994; Sugden, 1995; Janssen, 1998) and in the form of computer simulations (Kraus et al., 2000), strongly support Schelling's early intuition that choice processes exist that can facilitate people's coordination through the selection of the so-called *focal points*. The research agenda devised by Schelling then, consists in identifying "rational rules" accounting for the ability shown by humans to select focal points, and consequently coordinate, in the complete absence of communication.

We shall discuss in fuller detail the close connection between pure coordination games and Rationality-as-conformity later on in section 8.4, where it will be argued that Rationality-as-conformity does provide a solution concept for a certain class of coordination problems, as captured by the *conformity game*.

2.4 Social choice functions

Both the conventional theory of rational choice and the conventional theory of (non-cooperative) games recalled above are driven by an outcome-based characterisation of rationality. In particular we have seen that the former begins with imposing some consistency requirements on the individual's preferences. Those constraints then, lead to a representation of preference by means of some utility function, and against this background, the maximisation of expected utility is defined as the desired outcome of rational choices.

One important aspect of the outcome-based approach, which we haven't yet emphasised, is that those consistency requirements are imposed on the preferences defined on a set of options (actions in A) which is fixed once and for all. In contrast to

this, a key feature of the approach to rational choice that we are about to discuss is the characterisation of consistency constraints on the agents' choices across *varying* sets of possible options.

The traditional framework of *social choice theory* can surely be taken as representative of this approach. While economic rationality (and in particular game theory) aims at providing a normative framework for the rational (strategic) *interaction* among agents, it still focuses on decidedly "individualistic" features of decision making. Each agent aims only at maximising her own personal interests. The theory of social choice, on the other hand, extends this dimension by modelling the choice behaviour of the *homo sociologicus* whose aim is, to paraphrase Arrow, to mediate between individual values and social welfare. The fact that this idea of rationality must go beyond the pursuit of self interest is effectively schematised by Hammond:

> In my view, social choice theory should be about specifying suitable objectives for public officials and others responsible for major decisions affecting large numbers of individuals. (Hammond, 1997)

In fact we can consider social choice theory as a somewhat two-fold approach: with the first it aims at defining the conditions under which a "rational" aggregation of individual preferences (or judgments) turns out to be (im)possible. A notable example of this situation, which in many ways overlaps with the theory of *cooperative* games, is given by *voting systems*. The second approach, on the other hand, puts the emphasis on the definition of suitable conditions that any "rational" social choice process should satisfy. Of course the fact that these are just two sides of the same coin becomes clearly apparent by taking a social choice process to aim exactly at the aggregation of preferences (or judgments).

The main goal of the choice-function approach – the one on which we focus here – is to provide a set of constraints leading to the representation of rational choice in terms of "selection of the best options":

> The fulfilling of these [constraints] for a choice [function] is equivalent to

> the existence of its optimizational representation. (Aizerman and Malishevski, 1981, p.1030)

It is clear then, that as far as the characterisation of rational choice is concerned, what distinguishes the choice-functions approach from the one endorsed by the conventional theory of rational choice illustrated above is that it leads to a definition of "optimization" by means of constraints (i.e. properties, axioms, desiderata, etc.) imposed on the *choice process* rather than in terms of the desired properties of the *final outcome*. Put in our terminology, whilst the conventional theory of rational choice is outcome-based, the choice-function approach to social choice is fundamentally process-based.

It is therefore hardly surprising that there are many points of convergence between the target and the intuitions of Rationality-as-conformity and those of social choice theory. The remainder of this section provides a rough outline of the choice-function approach highlighting those common features.

The two fundamental ingredients of the choice-function model are:

- A set $\mathbb{W}$ of possible worlds $f, g, \ldots$, whose non-empty subsets are denoted by $K_1, K_2, \ldots$;
- A choice function R defined on every non-empty $K \subseteq \mathbb{W}$ such that $\emptyset \neq R(K) \subseteq K$.

For present purposes we shall intuitively read $R(K)$ *as the set* $\{f \mid f \in K\}$ *of possible worlds that a rational agent has reason to prefer over all the other (distinct) possible worlds in* K. In the choice theoretic literature, the function R is otherwise called a *choice function* or a *selection function*, whereas the set $R(K)$ is often referred to as the *choice set* or the set of *best elements of* K.

Note that, in a stringent parallel with the characterisation of belief as probability via betting behaviour recalled above, the choice function $R(K)$ is intended as an abstract model of an agent's disposition to choose from K, rather than an empirical model of an agent's actual choice behaviour.

The conventional approach to characterizing what amounts, for a rational agent, to having a reason to prefer $R(K)$ over its complement in K, pivots on imposing suitable constraints on the choice function R. The immense literature on the subject contains a conspicuous number of such principles (i.e. constraints). Some fundamental results in the area, however, invest the following with the status of being "core properties":

Property α (alias **Heritage**; **Independence of irrelevant alternatives**; **Coherence**) :

$$\text{if } K_1 \subseteq K_2, \text{ then } K_1 \cap R(K_2) \subseteq R(K_1).$$

Arrow's Axiom (alias **Strict Heritage**; **Weak Axiom Of Revealed Preference**):

$$\text{if } K_1 \subseteq K_2, \text{ and } R(K_2) \cap K_1 \neq \emptyset \text{ then } R(K_2) \cap K_1 = R(K_1).$$

Property γ (alias **Concordance; Expansion**):

$$R(K_1) \cap R(K_2) \subseteq R(K_1 \cup K_2).$$

Independence of rejecting the outcast variants (alias **Nash's axiom**):

$$\text{if } R(K_2) \subseteq K_1 \subseteq K_2 \text{ then } R(K_1) = R(K_2).$$

(Note that if the latter equality is replaced by the subset relation, then we obtain **Aizerman's axiom**.)

The above are usually considered to embed the core properties that any rational (social) choice function should satisfy as a consequence of a number of theorems, such as the Aizerman-Malishevski characterisation of *choice mechanisms*. Stated very informally, Theorem 1 of Aizerman and Malishevski (1981) asserts that certain combinations of the above properties are necessary and sufficient to represent rational choice functions in terms of the choice of the "best elements" from a given set of options K. A prominent, specific, consequence of this is that Property α is proven to be necessary and sufficient to ensure the *rationalizability* of a choice set by means of a simple ordering. We shall come back to this sort of characterisation of rational choice, which is not beyond internal criticism (see, e.g. Luce and Raiffa, 1957; Kalai et al., 2002), after the formal framework of Rationality-as-conformity will be introduced.

We note in passing that the above properties give rise in a natural way to a semantical characterization of *rational consequence relations* (Lehmann, 2001), and so are importantly connected to the study of nonmonotonic logics. (Rott, 2001, see esp. p. 153–163) reconstructs the original representation results for rationalizable choice functions and relates them to the study of nonmonotonic reasoning and belief revision.

2.5 Further remarks

The outcome-based vs. process-based distinction is surely not novel in the discussion of rational choice. A version of it can be traced back to the work of Simon, who however, puts the emphasis on somewhat different aspects of this distinction. In particular he stresses the fact that the outcome-based characterisation is sufficient for adequate prediction only under the assumptions usually endorsed within the economic-rationality approach. It becomes insufficient, however, once we take into considerations agents with *bounded rationality*, both in terms of knowledge and in terms of computational power. In this case, the process-based approach is needed:

> The rational person of neoclassical economics always reaches the decision that is objectively, or substantially, best in terms of the given utility function. The rational person of cognitive psychology goes about making his or her decisions in a way that is procedurally reasonable in the light of the available knowledge and means of computation. (Simon, 1986, p.211)

Among many others, Simon pioneered alternative accounts of the distinguishability among options of a choice problem, that is alternative to the comparison based on plain ordinal utility. A typical example, mentioned e.g. in Simon (1986) concerns people's decision as to whether insure against flood damage or not. If utility was the only yardstick, then all those agents for whom the "reimbursable damage from floods was greater than the premium" should buy the insurance. But this is in plain contradiction with the actual buyers of such a kind of insurance, typically individuals who

have been directly or indirectly involved into such events. If we want to understand how people behave when it comes to choosing whether to buy or not an insurance, we need to understand what makes it relevant for them to have one or not. And utility maximisation is clearly neither necessary nor sufficient to this end.

Chapter 3

Formalising Rationality-as-conformity

Abstract: *We introduce the key concepts and definitions intervening in the formalisation of the Rationality-as-conformity problem.*

The general pattern in the mathematical modelling of rational choice which has been outlined in the previous chapters consists of essentially two steps. The main goal of the first one is to provide a formalization of the mathematical *structure* within which the main problem is represented. This structure contains an explicit definition of all the features that are considered to be relevant to the formal statement and solution of the problem addressed by the corresponding "theory". The next step, then consists in proving the main results leading to an adequate *characterisation* of "rational choice". The formalization of Rationality-as-conformity constitutes no exception to this general pattern. Pivoting on the intuitions and the assumptions described in chapter 1.3 we shall:

- define an agent's *choice process*;
- formalise the Rationality-as-conformity *choice context*;
- formalise the main *choice problem* of Rationality-as-conformity.

This will provide all the necessary formal background to investigate the characterisations of Rationality-as-conformity given in chapters 4 –6.

3.1 Choice processes

The notion of a choice process to be introduced is strictly related to the Aizerman-Malishevski notion of a *choice mechanism*. As they point out (Aizerman and Malishevski, 1981, p.1031) a choice mechanism is specified by fixing an appropriately formalised *structure* on a given universe, together with a *rule* for selecting, given any non-empty subset of the universe, the set of its "best elements".

The present notion of a choice process, which reflects this two-folded nature, can be seen as a generalization of the Paris-Vencovská notion of an *inference process*. In a nutshell (see ch.6 of Paris, 1994, for precise details), an inference process is defined, for a consistent probabilistic knowledge base K, as an assignment of probability values to the sentences of SL such that, the values assigned are consistent with K. As recalled in chapter 1.2, the structure of the Paris-Vencovská characterisation is given in terms of a rich probability logic. The main problem there is to select a consistent solution to an agent's knowledge base, whereas the "rule" for choosing the solution is arrived at by imposing common sense constraints on the inference process itself.

In the formalization to follow, we will specify a much less structured universe, or *choice context*, than the one occurring in the Paris-Vencovská model. Yet, its spirit will be fully preserved as the set of "best options" is going to be specified by rules, or *Reasons*, for discarding those options which are inadequate for the purpose of achieving conformity.

A final important feature that the notion of a choice process inherits, so to speak, from that of an inference process is that it will be fully identified with an agent's reasoning about the specific choice context. In other words, in our idealised formal model, choice processes are taken to comprise the whole of the agent's "reasoning". This allows us to make precise the Likemindedness assumption illustrated above.

3.1.1 The choice context: possible worlds

The formalization of the main problem – choosing the same options we expect another like-minded yet non-communicating agent to choose – must begin by fixing the nature of the options that agents have to choose from. In line with the traditional terminology of (logical approaches to) uncertain reasoning, we shall refer to the options as *possible worlds*.

We denote the set of possible worlds $f, g, h \ldots$ (possibly with decorations) by $\mathbb{W}$, whilst non-empty subsets of $\mathbb{W}$ will be denoted by $K_1, K_2, \ldots$.

A set K is intuitively interpreted as an agent's *knowledge*, namely the knowledge that K contains all the possible options from which she must choose only one. Since this is clearly true for both the agents involved in a conformity problem, what this amounts to saying is that the set K contains *the* world on which the choice of each agent should converge. Notice that an immediate consequence of this conception is that the size of K introduces, if implicitly, a qualitative measure of uncertainty: the bigger the size of K, the greater the agents' uncertainty about which possible world will result in conformity. We can anticipate at this point that the characterisation of rationality pursued here aims at exploiting choice processes to reduce this uncertainty by the highest possible factor. This corresponds to a very general and fundamental idea in the logical approaches to reasoning under uncertainty (see Halpern, 2003, ch.2).

What is arguably the simplest possible choice context consistent with Rationality-as-conformity, is the one in which we have some finite non-empty set K of otherwise entirely structure-less options f, that is possible worlds that whilst different are otherwise entirely indistinguishable. Then the very definition of 'indistinguishable' seems to suggest that in this case there is no better strategy available to the agents involved in a conformity problem than to make a choice from K entirely at random – that is to say according to the uniform distribution.

It would seem natural to refer to this particular choice context as the *trivial* one.

The interesting thing to notice about it is that, in the trivial choice context, the "formalization" of a rational choice process is itself completely trivial. That is, assuming that the agents cannot refuse to choose, the only reasonable choice process that they could endorse is the one that makes no distinctions among the various options. In the light of our Fundamental assumption this amount to saying that in trivial contexts, there is nothing that rationality can do for the agents. Note that whilst the assumption that agents cannot refuse to choose is completely standard in the mathematical characterisations of rational choice, there are some recent investigations on the subject that drop this requirement, as in e.g. Rott (2001).

Of course we can, and perhaps should, expect more realistic choice contexts to be other than trivial. Mathematically, this means assuming that there indeed *is* some structure on the possible worlds. In particular there are properties which may or may not be true (to various degrees) of (or in) a world, the sum of which identifies the world uniquely. Examples of "possible worlds" in this sense are given by the Carnapian notion of *state description* or, in the propositional context, *atoms* as discussed in Paris (1994).

In the most general case, we may model possible worlds by taking $\mathbb{W}$ to be B^A, the set of all functions from a non-empty set A (of otherwise indistinguishable elements) to a non empty set B. As a consequence of this we shall sometimes denote the set of all non-empty subsets of $\mathbb{W}$ by $\wp^+(B^A)$. In fact our set of possible worlds is analogous to the set of acts $\mathbb{A}$ of Savage's framework (see section 2.2 above). However it is worth emphasising that we do not assume any structure whatsoever on the sets A and B.

Clearly the choice of the sets A and – more importantly – B may suggest itself certain "natural" interpretations of the set of possible worlds. For example taking A to be a set of propositional variables, thereby identifying A with a given propositional language L, various semantic interpretations can be naturally introduced by choosing appropriately the set B. So if B is say, the binary set $2 = \{0, 1\}$, the "natural" interpretation of worlds would be the classical, two-valued semantics. If, instead, we take the unit interval $[0, 1] = B$, this would "naturally" suggest taking worlds to define either degrees of belief or degrees of truth on the sentences formed from L. Or,

in the spirit of *plausibility measures*, we could just fix some minimal structure (i.e. a partial order) on an arbitrary set B and constrain the maps from A to B so as to satisfy certain desirable properties.

Rather than achieving full generality however, our intention in this work is to keep things as simple as possible whilst deferring the study of the other interesting cases to further investigations. As logicians then, the most obvious minimal structure on these possible worlds is that there are some finite number of unary predicates which each of them may or may not satisfy. To simplify matters for the present we shall further assume that each world is uniquely determined by the predicates it does or does not satisfy. In other words we are moving up from the language of equality of the trivial context to a finite unary language. What this amounts to then is that K is a non-empty subset of 2^A, the set of maps f from the finite non-empty set A into $2 = \{0, 1\}$.

3.1.2 Reasons

We understand Reasons are devices that agents apply to restrict their options, to go part, or sometimes even all, of the way to choosing a course of action or making a decision. In the present context, the terminology is motivated by the fact that those are the tools that agents utilize to *distinguish* among possible worlds. In consonance with our Fundamental assumption, this step gives agents *reasons* for their choices. As a typographical convention, we shall write "Reason" (capitalised) when referring to the specific component of a choice process, leaving the lower case for the informal meaning.

Formally, Reasons amount to choice (or selection) functions defined on the choice context described above (i.e. on non-empty sets of possible worlds). That is to say, functions $R : \wp^+(2^A) \longrightarrow \wp^+(2^A)$, where $\wp^+(2^A)$ is as usual the set of non-empty subsets of 2^A, will be called a *Reason* if

$$R(K) \subseteq K, \quad \forall K \in \wp^+(2^A).$$

It is immediate to realise that an *optimal* Reason is one that always returns a

singleton $R(K)$ for every $K \in \wp^+(2^A)$. Such a Reason would be optimal to the extent that its adoption would entail conformity with probability 1. We shall see, however, that this situation represents the exception rather than the rule in the formalization to follow. In the epistemic interpretation attached to the elements of $\wp^+(2^A)$, it could be said that whenever $|R(K)| = 1$, then the agents' choice process eliminated all the uncertainty present in their knowledge. Hence the sub-optimality of Reasons amount to nothing but the fact that in the Rationality-as-conformity framework we are dealing with *uncertain reasoning* in what is perhaps one of its most general and basic forms.

At the opposite extreme of the spectrum we locate the *trivial* Reason, that is to say the Reason R satisfying

$$R(K) = K, \quad \forall K \in \wp^+(2^A). \tag{3.1}$$

In practice, the progress ensured by the trivial Reason amounts to nothing at all! Its interpretation, in the present context, is that the options among which agents have to choose are too undistinguished for them to make, in accordance to the Fundamental assumption, a reasoned choice. In other words, the choice from the set K at hand is simply too hard.

Far more interesting and realistic, as we shall shortly see, is the case of non trivial (yet often sub-optimal) Reasons. If $|R(K)| > 1$ we seem to have two possible ways of proceeding. In the former we consider the possible worlds in $R(K)$ as all "equally good" to the agents' lights, so in accordance to our Fundamental assumption, we let the agent finalize the choice by picking one element from $R(K)$ according to the uniform distribution. This solution rests on the underlying assumption that the options are so structure-less yet distinct that it is not possible for them to compromise and select any given combination of them. The second solution is to combine in some appropriate way the possible worlds in $R(K)$ so as to obtain a singleton choice. Consider, for example, the case in which $R(K) = \{f, g, h\}$. The idea here would be define an appropriate concatenation $*$ on the set of possible worlds, allowing us to reduce the above to the singleton $R(K) = \{f * g * h\}$.

Those two strategies – randomizing or combining among the best options – bear an interesting parallel to two established approaches to non-monotonic reasoning, usually termed as *bold* (or credulous) and *cautious* (or sceptical), respectively. In this area the latter strategy is often preferred (see, e.g. Rott, 2001, 146). Interestingly, however, we shall find out that Rationality-as-conformity goes some way towards combining these two perspectives in the study of the Minimum Ambiguity Reason introduced in chapter 5 below.

Finally a word of notational conventions. Since we shall keep the choice context fixed throughout we shall refer to Reasons and choice processes interchangeably and we shall sometimes denote $\wp^+(2^A)$ as $\mathbb{K}$.

3.2 The main problem formalised

In order to illustrate the formalization of the Rationality-as-conformity main choice problem let's fix $A = 4$ and let K be the set of functions (possible worlds) $\{f_1, f_2, f_3, f_4, f_5\}$ where

	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.

In this case the conformity problem, for a pair of like-minded yet inaccessible agents, amounts to selecting one of the above rows so as to agree with each others' choice. However in presenting the problem like this we should be aware that as far as the agents are concerned there is not supposed to be any structure on A. Hence there is no further structure on K beyond the fact that it is the (unordered) set $\{f_1, f_2, f_3, f_4, f_5\}$. For practical examples this can be accomplished by informing the first agent that his or her counterpart may receive the matrix

0	0	0	1
0	1	0	0
0	1	1	0
1	1	1	1
0	0	1	0

Figure 3.1: A matrix representation of K

with the columns permuted and the rows possibly permuted. That is to say, in place of the matrix represented in figure 3.1, the agents might equally have, say either of the matrices represented in figure 3.2 (where compared to matrix in figure 3.1 the one on the left has the "second" and "third" rows permuted, whilst the one on the right has the "first" and "second": columns permuted):

0	0	0	1
0	1	1	0
0	1	0	0
1	1	1	1
0	0	1	0

0	0	0	1
1	0	0	0
1	0	1	0
1	1	1	1
0	0	1	0

Figure 3.2: Alternative matrix representations of K

The main goal of the Rationality-as-conformity framework can thus be rephrased as the characterisation of the *choice processes* which, if adopted by the agents, would enable (or at least facilitate) conformity. In fact we shall introduce three distinct choice processes as possible *solutions* for the Rationality-as-conformity problem. The common pattern among them can be illustrated again in analogy with the Aizerman-Malishevski notion of a choice mechanism.

In their discussion of a *General theory of best variants choice* (Aizerman and Malishevski, 1981), they define several mechanisms for distinguishing among "variants" or, in our terminology, possible worlds. The most basic one – the *scalar optimization choice* (otherwise known as the *rationalization of choice functions*) – is simply specified by fixing a certain choice context, K in our case, and a mapping, say ϕ from the

set of possible worlds to the "worse-better axis". The key aspect of this choice mechanism emerges when one takes its associated selection "rule" to be the *maximisation* of the value of ϕ (on its restriction to the choice context K). More precisely, the scalar optimization choice mechanism, when applied to the choice context K returns the set:

$$R(K) = \{f \in K \mid \phi(f) \geq \phi(g), \forall g \in K\}. \tag{3.2}$$

In the light of this analogy our notion of a Reason can be interpreted as an act of maximisation by means of which agents select the "best elements" from a given choice context K. As we shall see, the structure of the Rationality-as-conformity problem allows us to construct and justify three distinct ways of identifying the best elements relative to a given choice context.

One might question at this point whether a better model for the agent's actions might be to have him or her put a probability distribution over K and then pick according to that distribution. In fact in such a case the agent would do at least as well by instead selecting the most probable elements of K according to this distribution and then randomly (i.e. according to the uniform distribution) selecting from them – which puts us back into the original situation.

Chapter 4

The Regulative Reasons

ABSTRACT: *Our first characterisation of a choice process facilitating conformity is based on the adherence of reasons to three "common sense principles" which generalize analogous principles investigated in the Paris-Vencovská probabilistic logic. Hence we call the corresponding Reasons "Regulative".*

As illustrated in section 1.2 this work was in part motivated by considering why the principles of probabilistic uncertain reasoning introduced in the Paris-Vencovská characterisation warranted the description 'common sense'. Recall that the underlying problem in such a framework is entirely analogous to the one we are considering here, that is how to sensibly choose one out of a set of probability functions. It has been emphasised as well that the Paris-Vencovská solution is essentially process-driven as it requires that the choice process should satisfy such common sense principles. In fact this turns out well in the linear cases considered in Paris and Vencovská (1990, 1997) as the imposed principles happily permitted only one possible choice, as illustrated by Theorem 1.1 above.

Given such a fortunate outcome there, it would seem natural to attempt a similar procedure here, namely to specify certain 'common sense' principles we would wish the agent's Reasons to satisfy and see what comes out. Clearly, the present problem is much less structured then the one in which belief is represented via subjective probability functions. Indeed the current setting is arguably one of the simplest

ones in which we can make sense of rational choice concerning "knowledge" and "possibilities". It therefore follows that if choice processes analogous to the ones that characterise probabilistic common sense could be specified, those would have an undoubtedly high level of generality.

Our next step then is to introduce 'common sense principles' or rules that, arguably, Reasons *should* satisfy if they are to prevent agents from undertaking "unreasonable steps". Hence, we call the resulting Reason(s), *Regulative*. The key result of this section is that their observance leads to a characterisation of a set $R(K)$ of "naturally outstanding elements" of K, formulated in Theorem 4.2.

4.1 Regulative Reasons defined

We now introduce the common sense principles constraining our first choice process for Rationality-as-conformity. As they all generalize some principle of probabilistic uncertain reasoning of the Paris-Vencovská characterisation introduced informally in section 1.2 above, we preserve here the original names. For the sake of keeping the presentation compact, we defer the discussion of these principles to the next section.

Renaming:

> Let $K \in \mathbb{K}$ and let σ be a permutation of A. R satisfies *Renaming* if whenever $K\sigma = \{ f\sigma \mid f \in K \}$ then
>
> $$R(K\sigma) = R(K)\sigma.$$

(Notice that $R(K)\sigma$ is, as usual, the set $\{ f\sigma \mid f \in R(K) \}$.)

Intuitively, Renaming captures the idea that inessential changes in the presentation of the choice problem (that is to say in the representation of the matrix K) should not introduce grounds for distinguishing among possible worlds in K.

The justification for Renaming depends essentially on the formal properties of the choice context defined above. In particular, since the elements of A have no further structure other than being a set of distinct elements, any permutation of

these elements simply produces an exact replica of what we started with. Hence, such a permutation should be completely irrelevant in terms of *distinguishing among options*: if an agent feels that the "best choice" of worlds from K coincides with the set of worlds $R(K)$ then she should feel the same for these replicas, i.e. that the "best choices" of worlds from $K\sigma$ should be $R(K)\sigma$.

Obstinacy:

Let $K_1, K_2 \in \mathbb{K}$. R satisfies *Obstinacy* if whenever $R(K_1) \cap K_2 \neq \emptyset$ then

$$R(K_1 \cap K_2) = R(K_1) \cap K_2.$$

The rationale for this principle is that if each agent expects the other's choices from K_1 to be $R(K_1)$ where in fact some of these possible worlds are also elements of K_2 then such worlds must still remain the "best elements" were the choice to be restricted to $K_1 \cap K_2$. The refinement of K_1 to $K_1 \cap K_2$ gives no reason to the agents to turn any of the non-preferred element of K_1 into a preferred one. On the other hand, the preferred options from $K_1 \cap K_2$ should all be included among the preferred options from K_1 which happen to be also in K_2. We shall illustrate this further with some examples in section 4.1.1 below.

In order to introduce our final principle we need a little notation. For $K \in \wp^+(2^A)$ we say that $X \subseteq A$ is a *support* of K if whenever $f, g \in \mathbb{W}$ and $f \restriction X = g \restriction X$ then

$$f \in K \iff g \in K.$$

The set A itself is trivially a support for every $K \in \wp^+(2^A)$. More significantly we can show that every $K \in \wp^+(2^A)$ has a unique smallest support.

Lemma 4.1. *If $K \in \wp^+(2^A)$ has a support, than there is a unique smallest finite support X for K.*

Proof. We have to show that the support property is closed under intersection, so let X_1, X_2 be support for K and let $Y = X_1 \cap X_2$. To show that Y is a support for K it is enough to show that for $f, g \in \mathbb{W}$ if $f \restriction Y = g \restriction Y$ then $f \in K \Leftrightarrow g \in K$. For

given such f, g, Y we can pick from $\mathbb{W}$ a function h which agrees with f on $X_1 - Y$ and with g on $X_2 - Y$. Therefore:

$$f \in K \iff h \in K$$
$$\iff g \in K.$$

■

Notice that if K has support X then $K\sigma$ has support $\sigma^{-1}X$. If K has support X then it is useful to think of this knowledge as telling the agent (just) how elements of K act on X. Namely, for f to be in K it is necessary and sufficient that $f \restriction X = g$ for some

$$g \in \{\, h \restriction X \mid h \in K \,\}.$$

We can now formulate the following principle.

Irrelevance:

Suppose $K_1, K_2 \in \mathbb{K}$ are such that $K_1 \cap K_2 \neq \emptyset$ and have supports X_1, X_2 respectively. If for any $f_1 \in K_1$ and $f_2 \in K_2$ there exists $f_3 \in \mathbb{W}$ such that $f_3 \restriction X_1 = f_1 \restriction X_1$ and $f_3 \restriction X_2 = f_2 \restriction X_2$ then,

$$R(K_1) \restriction X_1 = R(K_1 \cap K_2) \restriction X_1$$

where

$$R(K) \restriction X = \{\, f \restriction X \mid f \in R(K)\}.$$

The condition on K_1, K_2 amounts to saying that *as far as* K_1 *is concerned* K_2 *is irrelevant (and conversely)* because given that we know (only) that f satisfies the requirement for membership of K_1 (i.e. that $f \restriction X_1$ is amongst some particular set of functions on X_1) the additional information that $f \in K_2$ tells us nothing we didn't already know about $f \restriction X_1$.

The principle then amounts to saying that in these circumstances the choices from K_1 and K_2 should also reflect that irrelevance. That is, if $f_1 \in R(K_1)$, then there is an $f_3 \in R(K_1 \cap K_2)$ such that $f_3 \restriction X_1 = f_1 \restriction X_1$ and conversely given $f_3 \in R(K_1 \cap K_2)$ there exists such a f_1 (and similarly for K_2).

Definition. *We shall say that a Reason R is a* Regulative Reason*, if is satisfies Renaming, Obstinacy and Irrelevance.*

4.1.1 Comments on the principles

Renaming

A widely influential property of probabilistic uncertain reasoning related to Renaming is de Finetti's notion of *Exchangeability*. Suppose we are considering a series of observations (say draws from an urn) to which we can attach certain properties. Exchangeability then is satisfied whenever the inference based on such observations is independent of their ordering. (See de Finetti (1995) and Kuipers (1998) p.533)

In the social choice literature, on the other hand, an entirely analogous property to Renaming is assumed, namely the axiom of *Anonymity*. In a nutshell this amounts to requiring that a social aggregation function (i.e. a function aggregating the individual preferences or judgments of a whole society into one single preference or judgment) should be invariant under permutations of the individuals' "names". Anonymity is one of the properties assumed by May in his characterisation of simple majority vote (May, 1952). Note also that an entirely analogous assumption to Renaming – *Symmetry* – is made by Nash in his solution to the bargaining problem (Nash, 1950).

Obstinacy

The 'justification' we proposed for the acceptance of Obstinacy as a commonsensical principle is, *in general*, more than a little suspect. For instance, consider the case in which by intersecting K_1 with K_2 some otherwise rather nondescript world from K_1 becomes, within $K_1 \cap K_2$, sufficiently distinguished to be a natural choice. Whilst this will become clearer later when we have other Reasons to hand, it can nevertheless still be illustrated informally at this point.

Example 7. Suppose that K is

$$\begin{matrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{matrix}$$

and let suppose further that the best elements from this set were $R(K) = \{0000, 1111\}$.

Now if we take $K' \subset K$ to be

$$\begin{matrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{matrix}$$

then it one might argue that $R(K') = \{1100\}$ has now become the obvious choice, not (a random element of) $\{0000, 1111\}$, thus contradicting Obstinacy. Reaching this conclusion, however, requires *changing* the way we look at options, or equivalently, adopting distinct choice processes for the same choice problem. Yet if we take agents to be like-minded and "stable" in their adoption of Reasons for the solution of any given choice problem, then the occurrence of this phenomenon is ruled out. (See 5.4 for more on this.)

An analogous point could be raised by considering the following example, a familiar objection raised against the Maximum Entropy Inference Process and in favour of the Centre of Mass Inference Process (see e.g. Jaeger, 1998).

Example 8. Suppose that agents are choosing points from closed subintervals of $[0, 1]$. Some appeal to symmetry might convince them that *the* "natural" choice from $[0, 1]$ itself should be $1/2$. However, thoughts along similar lines might suggest $1/4$ for the choice from $[0, 1/2]$ (because $1/4$ has suddenly become 'distinguished' when we intersected $[0, 1]$ with $[0, 1/2]$) when Obstinacy would point instead to the choice of $1/2$.

Despite these potential difficulties, we still feel that in the context of our investigations it is of some theoretical interest at least to persevere with this principle, and also because of the conclusions it leads to. Moreover, in so as far as nothing specific is assumed about the nature of the options, the property captured by Obstinacy is widely endorsed by the social choice community under the form of the *Independence of the irrelevant alternatives* (Independence, for short).

Independence, which we already encountered in section 2.4 above, is a key, if highly debated, principle of the choice-theoretic characterisations of rationality. Intuitively, it states that the "best options" of a given set remain "best" in every subset containing them. Hence it is justified as a principle warranting the internal consistency of choice across varying menus and it turns out to be the key property ensuring the *rationalizability* of choice functions (see section 2.4). In fact, if the rationalizing relation is a weak ordering then the axiom is a necessary and sufficient condition for the representation of choice by means of such an ordering (Sen, 1986). The importance of this result is often emphasised by referring to Independence as the "rationality axiom" (Kalai et al., 2002).

In the social choice literature, the general criticisms towards Independence and hence, by analogy, to the requirement imposed on rational choice by Obstinacy, depend on the assumptions concerning some form of background knowledge. Thus, for example, Sen (1997) points out that an agent i might choose an invitation to have a cup of tea at j's place in preference to going straight home after work, yet choose to rush home if j offers cocaine and heroin with the tea. Sen's argument, which goes along the lines of the so-called "Luce and Raiffa dinner" (see Luce and Raiffa, 1957), is that the two sets of options have for i distinct *epistemic values*. Specifically, the latter, enriched set of options would allow i to draw tentative, plausible, etc., conclusions about the particular situation at hand which would not be possible in the former. So for instance i might use defaults like "usually decent people do not make use of heroin" to conclude that "it's quite likely that j is not a decent person". This, together with i's belief that "having tea is enjoyable only with decent people" may lead i to prefer one course of action over the other.

According to Kalai et al. (2002), i's violation of the Independence axiom need not lead to irrational choice. Their proposal is in fact to account for i's rationality by explaining its choice behaviour by means of *multiple rationales*, that is to say by defining the standard maximization procedure on a (finite) number of preference relations defined the choice context instead of just a single ordering. Hence, different sets of options might support different rationalization orderings.

An interesting case of the rationalization by multiple rationales is the so-called (u, v) procedure discussed in Kalai et al. (2002). In a nutshell, it recommends that an agent i maximizes a utility function u so long as the expected value of the maximization of u is above a distinguished value $\mathbf{v}$ of the utility function v. Should the maximal value of u fall below $\mathbf{v}$, then i should pick the $v-$maximal element. A situation in which it seems realistic that agents could (or should) adopt a (u, v) procedure arises from taking u to be "social welfare" and v "personal welfare". Then, an agent maximises the social welfare so long as doing so does not threaten her own well-being.

Note that the distinguished $v-$value $\mathbf{v}$ might roughly be taken as a measure of i's altruism. In the domain of artificial intelligence, the (u, v) procedure might be interpreted in terms of planning: take u to represent the "main utility" relative, that is to the agent's current goals and v to represents some sort of "background utility", relative that is to, say, the avoidance of obstacles.

We conclude this small digression on Obstinacy and related principles by noting that while all those potential criticisms may well be grounded on specific limitations of the property captured by Obstinacy, we feel that the formalization within which we are considering our choice problem is so abstract that most of the points mentioned above would fail to have a direct bearing on the main Rationality-as-conformity problem.

Irrelevance

The justification for Irrelevance goes along the following lines. In choosing a "best element" from K_1 agents are effectively choosing from $K_1 \restriction X_1$ and then choosing

from all possible extensions (in $\mathbb{W}$) of these maps to domain A, and similarly for K_2. The given conditions allow that in choosing from $K_1 \cap K_2$ agents can first freely choose from $K_1 \restriction X_1$ then from $K_2 \restriction X_2$ and finally freely choose from all possible extensions to domain A. Viewed in this way it seems then that any function in $R(K_1) \restriction X_1$ should also be represented in $R(K_1 \cap K_2) \restriction X_1$. Notice, that there seems to be an implicit assumption in this argument that for $f \in K_1$, $f \restriction X_1$ and $f \restriction A - X_1$ are somewhat independent of each other. In the current simple case of $\mathbb{W} = 2^A$ this is true but it fails in the case, considered in section 7.1 below, in which the worlds are probability functions.

4.2 Regulative Reasons characterised

We start by noticing that there certainly is one Reason satisfying the common sense properties defined above, namely the *trivial Reason* R such that $R(K) = K$ for all $K \in \mathbb{K}$, though of course in practice this 'reason' amounts to nothing at all. It turns out that if we had taken A to be infinite and $\mathbb{K}$ the non-empty subsets of 2^A with finite support (so $R : \mathbb{K} \longrightarrow \mathbb{K}$) then the trivial one would have been the only Regulative Reason, as shown by Proposition 4.12 below.

Theorem 4.2. *Let R be a Regulative Reason. Then either R is trivial or $R = R_0$ or R_1 where for $i = 0, 1$ R_i is defined by*

$$R_i(K) = \{\, f \in K \mid \forall\, g \in K, |f^{-1}(i)| \geq |g^{-1}(i)| \,\}.$$

Conversely each of these three Reasons are Regulative, i.e. satisfy Renaming, Obstinacy and Irrelevance.

We begin with the proof of the "if" part. As usual, $\vec{0} : A \longrightarrow 2$ is defined by $\vec{0}(x) = 0$ for all $x \in A$ and similarly, $\vec{1} : A \longrightarrow 2$ is defined by $\vec{1}(x) = 1$ for all $x \in A$.

The first step consists in showing that Regulative Reasons are indeed three-fold.

Lemma 4.3. *Let R be Regulative. Then either $R(2^A) = 2^A$ or $R(2^A) = \left\{\vec{0}\right\}$ or $R(2^A) = \left\{\vec{1}\right\}$.*

Proof.

We first show the following claim:

If $f, g \in R(2^A)$ (possibly $f = g$) are such that $0, 1$ are in the ranges of f, g respectively, then $R(2^A) = 2^A$.

To this end let $f, g \in R(2^A)$ and $f(x) = 0$ and $g(y) = 1$ for some $x, y \in A$. For σ a permutation on A transposing only x and y we have that $2^A\sigma = 2^A$. Hence, by Renaming, $R(2^A)\sigma = R(2^A\sigma)$. In particular:

$$f \in R(2^A) \Rightarrow f\sigma \in R(2^A). \tag{4.1}$$

Now let $K = \{h \in 2^A \mid h(y) = 0\}$. Since $f\sigma \in R(2^A) \cap K \neq \emptyset$ then:

$$\begin{aligned} R(2^A) \cap K &= R(2^A \cap K) \quad \text{(by Obstinacy)} \\ &= R(K). \\ \therefore f\sigma &\in R(K). \end{aligned} \tag{4.2}$$

If we take $X_1 = A - \{y\}$ and $X_2 = \{y\}$ to be supports of 2^A and K respectively, we can see that since $\emptyset = \{y\} \cap X_1$, for any $f_1 \in 2^A$ and $f_2 \in K$, we can construct a function $f_3 \in 2^A$ such that $f_3 \restriction X_1 = f_1 \restriction X_1$ and $f_3 \restriction X_2 = f_2 \restriction X_2$. Thus

$$\begin{aligned} R(2^A) \restriction X_1 &= R(2^A \cap K) \restriction X_1 \quad \text{(by Irrelevance)} \\ &= R(K) \restriction X_1. \end{aligned} \tag{4.3}$$

Therefore, $g \restriction X_1 \in R(K) \restriction X_1$. Furthermore for

$$g'(z) = \begin{cases} g(z) & \text{if} \quad z \neq y \\ \\ 0 & \text{if} \quad z = y. \end{cases} \tag{4.4}$$

we have that $g' \in R(K)$. Hence $g' \in R(2^A)$, by (4.2) above.

The claim now follows since we have shown that if we take any function $h \in R(2^A)$ and change its value on one argument the resulting function is also in $R(2^A)$.

The proof of Lemma 4.3 now follows by noticing that if $R(2^A) \neq 2^A$ then by the claim either 0 or 1 is not in the range of any $f \in R(2^A)$. Therefore, since $R(2^A) \neq \emptyset$ it must either be that $R(2^A) = \{\vec{0}\}$ or $R(2^A) = \{\vec{1}\}$. ■

Our next step is to prove the required result for trivial Reasons.

Lemma 4.4. *If $R(2^A) = 2^A$, then $R(K) = K$ for any $K \in \mathbb{K}$.*

Proof. Notice that if $R(2^A) = 2^A$ then for $K \in \mathbb{K}$,

$$K \cap R(2^A) = K \neq \emptyset$$

so by Obstinacy,

$$R(K) = K \cap R(2^A) = K.$$

■

Hence, the final step in the proof of the "if" direction of Theorem 4.2 deals with the more interesting case of non-trivial Reasons.

It will be useful here to introduce a little notation. For the remainder of this section, let $\pi : dom(\pi) \longrightarrow \{0, 1\}$, where the domain of π, $dom(\pi)$, is a subset of A. Similarly for $\pi_1, \ldots, \pi_k$. For such a π let

$$K_\pi = \{\, f \in 2^A \mid f \restriction dom(\pi) = \pi \,\}.$$

Lemma 4.5. *If $R(2^A) = \left\{\vec{1}\right\}$, then*

$$R(K_\pi) = \left\{\pi \vee \vec{1}\right\},$$

where

$$\pi \vee \vec{1}(x) = \begin{cases} \pi(x) & \text{if } x \in dom(\pi) \\ \\ \vec{1}(x) & \text{otherwise.} \end{cases} \tag{4.5}$$

Proof. Suppose that $z \in A - dom(\pi)$. To prove the result it is enough to show that $f(z) = 1$ for $f \in R(K_\pi)$. Let $\{z\}$ and $dom(\pi)$ be supports of 2^A and K_π

respectively. Notice that the conditions for the applications of Irrelevance are met since $\emptyset = \{z\} \cap dom(\pi)$. Hence

$$R(2^A) \restriction \{z\} = R(K_\pi) \restriction \{z\}\,.$$

Therefore, for $f \in R(K_\pi)$

$$f(z) = \begin{cases} 1 & \text{if } z \in A - dom(\pi), \\ \\ \pi(x) & \text{if } z \in dom(\pi), \end{cases} \tag{4.6}$$

making $f = \pi \vee \vec{1}$.

■

This can be immediately generalized as follows.

Lemma 4.6. *Suppose* $R(2^A) = \left\{\vec{1}\right\}$ *and let* $Z = \{z_1, z_2, \ldots, z_n\} \subseteq A$ *with* $0 \leq r \leq n$. *Let* $\tau_1^r, \tau_2^r, \ldots, \tau_q^r$ *be all the maps from a subset of size* r *of* Z *to* $\{0\}$. *Then*

$$R(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}) = \left\{\tau_1^r \vee \vec{1}, \tau_2^r \vee \vec{1}, \ldots, \tau_q^r \vee \vec{1}\right\}.$$

Proof.

We first recall that, by the definition of R,

$$R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \subseteq \left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \tag{4.7}$$

Now let σ be a permutation of A such that $Z\sigma = Z$. Then

$$\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right)\pi = \left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right).$$

Hence, by Renaming:

$$f \in R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \iff f\sigma \in R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \tag{4.8}$$

By equation (4.7), $R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \cap K_{\tau_j^r} \neq 0$, for some $0 \leq j \leq q$. Thus, by Obstinacy,

$$\begin{aligned} R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \cap K_{\tau_j^r} &= R\left(\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \cap K_{\tau_j^r}\right) \\ &= R\left(K_{\tau_j^r}\right) \quad \text{(for some } 0 \leq j \leq q\text{)}. \end{aligned} \tag{4.9}$$

Recalling, from Lemma 4.5, that $R\left(K_{\tau_j^r}\right) = \left\{\tau_j^r \vee \vec{1}\right\}$ we have that $\tau_j^r \vee \vec{1} \in R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right)$ for some $0 \leq j \leq q$. By equation (4.8), however, this can be generalized to any $0 \leq j \leq q$. Hence

$$R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \supseteq \left\{\tau_1^r \vee \vec{1}, \tau_2^r \vee \vec{1}, \ldots, \tau_q^r \vee \vec{1}\right\}. \tag{4.10}$$

To see that the converse is also true, suppose $h \in R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right)$. Then since

$$R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \subseteq K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r},$$

$h \in K_{\tau_j^r}$, for some j. But as we have just observed,

$$R\left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}\right) \cap K_{\tau_j^r} = R\left(K_{\tau_j^r}\right),$$

so $h = \left\{\tau_j^r \vee \vec{1}\right\}$, as required. ■

Lemma 4.7. *Suppose $Z = \{z_1, z_2, \ldots, z_n\} \subseteq A$ and let $\tau_1^r, \tau_2^r, \ldots, \tau_p^r$ be some maps from a subset of Z of size r to $\{0\}$. Then*

$$R(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_p^r}) = \left\{\tau_1^r \vee \vec{1}, \tau_2^r \vee \vec{1}, \ldots, \tau_p^r \vee \vec{1}\right\}.$$

Proof. Let $\tau_1^r, \tau_2^r, \ldots, \tau_q^r$ be as in Lemma 4.6. Then by Obstinacy

$$\begin{aligned} R(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_p^r}) &= R(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_q^r}) \cap \left(K_{\tau_1^r} \cup K_{\tau_2^r} \cup \ldots \cup K_{\tau_p^r}\right) \\ &= \left\{\tau_1^r \vee \vec{1}, \tau_2^r \vee \vec{1}, \ldots, \tau_p^r \vee \vec{1}\right\}. \end{aligned}$$

■

We now have all the devices necessary to move on to the crucial step.

Lemma 4.8. *Let $\tau_1^{r_1}, \tau_2^{r_2}, \ldots, \tau_p^{r_p}$ be maps each from some subset of Z of cardinality $r_1, \ldots r_p$ to $\{0\}$ respectively. If $R(2^A) = \left\{\vec{1}\right\}$, then for $r = \min\{r_i \mid i = 1, \ldots, p\}$*

$$R\left(K_{\tau_1^{r_1}} \cup K_{\tau_2^{r_2}} \cup \ldots \cup K_{\tau_p^{r_p}}\right) = \left\{\tau_j^{r_j} \vee \vec{1} \mid r_j = r\right\}.$$

Proof. Let $\delta^r_1, \delta^r_2, \ldots, \delta^r_q$ be all the maps from a subset of size r of Z to $\{0\}$. Then

$$\left\{\tau_j^{r_j} \vee \vec{1} \mid r_j = r\right\} \subseteq \left\{\delta_i^r \vee \vec{1} \mid i = 1, \ldots, q\right\}. \tag{4.11}$$

Now, since each $K_{\tau_i^{r_i}} \subseteq K_{\delta_k^r}$, for some k, by Lemma 4.7 above and (4.11)

$$\begin{aligned} R\left(K_{\tau_1^{r_1}} \cup K_{\tau_2^{r_2}} \cup \ldots \cup K_{\tau_p^{r_p}}\right) &= R(K_{\delta_1^r} \cup K_{\delta_2^r} \cup \ldots \cup K_{\delta_q^r}) \cap \left(K_{\tau_1^{r_1}} \cup K_{\tau_2^{r_2}} \cup \ldots \cup K_{\tau_p^{r_p}}\right) \\ &= \left\{\tau_j^{r_j} \vee \vec{1} \mid r_j = r\right\}. \end{aligned}$$

■

Corollary 4.9. *For $K \in \wp^+(2^A)$, if $R(2^A) = \{\vec{1}\}$ then*

$$R(K) = \left\{f \in X \mid \left|f^{-1}\{0\}\right| = r\right\},$$

where r is minimal such that $|f^{-1}\{0\}| = r$ for some $f \in K$.

Proof. The result follows as an immediate consequence of Obstinacy and Lemma 4.8. ■

Notice that by duality, Corollary 4.9 holds for $\vec{1}$ being replaced by $\vec{0}$.

This completes the proof of the "if" direction of Theorem 4.2. We now move on to show its converse, namely that if a Reason $R(\cdot)$ is defined in any of the above three ways, then Renaming, Irrelevance and Obstinacy are satisfied. This clearly characterises completely Regulative Reasons for the special case in which worlds are maps from finite set A to 2.

Again, we start with the trivial Reasons, and then we move on to the case of the non-trivial ones.

Lemma 4.10. *Suppose $R(K) = K$, for all $K \in \wp^+(2^A)$. Then Renaming, Obstinacy and Irrelevance are satisfied.*

Proof. (Renaming) Suppose $K \in \wp^+(2^A)$ with support $X \subseteq A$ and σ is a permutation of A. Then

$$R(K)\sigma = K\sigma = R(K\sigma)$$

as required.

(Obstinacy) For $K_1, K_2 \in \wp^+(2^A)$, with supports $X_1, X_2 \subseteq A$ respectively,

$$R(K_1) \cap K_2 = K_1 \cap K_2 = R(K_1 \cap K_2)$$

as required.

(Irrelevance) Suppose $K_1, K_2 \in \wp^+(2^A)$ (with supports X_1, X_2 respectively) are such that for any $f_1 \in K_1, f_2 \in K_2$, there exists $f_3 \in \mathbb{W}$ such that $f_3 \restriction X_1 = f_1 \restriction X_1$ and $f_3 \restriction X_2 = f_2 \restriction X_2$. We have to show that $R(K_1) \restriction X_1 = R(K_1 \cap K_2) \restriction X_1$. Let $g \in 2^{X_1}$. If $g \in R(K_1 \cap K_2) \restriction X_1$ then obviously $g \in R(K_1) \restriction X_1$. As to the other direction, suppose $g = f_1 \restriction X_1$ with $f_1 \in K_1$. Then we are given that for $f_2 \in K_2$ there is $f_3 \in \mathbb{W}$ such that $f_3 \restriction X_1 = f_1 \restriction X_1 = g$ and $f_3 \restriction X_2 = f_2 \restriction X_2$ Thus, $f_3 \in K_1 \cap K_2$ and $g = f_3 \restriction X_1 \in R(K_1 \cap K_2) \restriction X_1$, as required. ■

Lemma 4.11. *$R_1(K)$ satisfies Renaming, Obstinacy and Irrelevance.*

Proof.

(Renaming) Let σ be a permutation of A. Then

$$\begin{aligned} f \in R_1(K)\sigma &\iff f = g\sigma, \text{for some } g \in R_1(K) \\ &\iff f = g\sigma, \text{for some } g \in \left\{h \in K \mid \left|h^{-1}\{1\}\right| = r\right\} \end{aligned} \tag{4.12}$$

where $r = \max\{|h^{-1}\{1\}| \mid h \in K\}$. But since $|h^{-1}\{1\}| = |(h\sigma)^{-1}\{1\}|$, then

$$h \in K \text{ and } \left|h^{-1}\{1\}\right| = r \iff h\sigma \in K\sigma \text{ and } \left|(h\sigma)^{-1}\{1\}\right| = r.$$

and $r = \max\{|(h\sigma)^{-1}\{1\}| \mid h\sigma \in K\sigma\}$. Hence

$$f \in R_1(X) \iff f\sigma \in R_1(X\sigma),$$

as required.

(Obstinacy) Let $K_1, K_2 \in \wp^+(2^A)$ and let $R_1(K_1) \cap K_2 \neq \emptyset$ and set

$$r' = \max\left\{\left|g^{-1}\{1\}\right| \mid g \in K_1 \cap K_2\right\}.$$

We claim that $r' = r$, where r is defined as above. To see that the result follows from this claim notice that if $r' = r$, then

$$\begin{aligned} R_1(K_1 \cap K_2) &= \left\{ f \in K_1 \cap K_2 \mid \left| f^{-1}\{1\} \right| = r \right\} \\ &= \left\{ f \in K_1 \mid \left| f^{-1}\{1\} \right| = r \right\} \cap K_2 \\ &= R_1(K_1) \cap K_2. \end{aligned}$$

We show the claim by contradiction. Since $K_1 \cap K_2 \subseteq K_1$, the case $r' > r$ is clearly not possible. To see that $r' < r$ is not possible either, and hence that $r' = r$, let $h \in R_1(K_1) \cap K_2$. Then r' would be the largest n for which there exists $h' \in K_1 \cap K_2$ such that $\left| h'^{-1}\{1\} \right| = n$. But since $h \in R_1(K_1)$, r would be such an n, giving $r' \geq r$ as required.

(Irrelevance) Suppose $K_1, K_2 \in \wp^+(2^A)$ (with supports X_1, X_2, respectively) and for any $f_1 \in K_1, f_2 \in K_2$, there exists $f_3 \in \mathbb{W}$ such that $f_3 \restriction X_1 = f_1 \restriction X_1$ and $f_3 \restriction X_2 = f_2 \restriction X_2$. We have to show that

$$R_1(K_1) \restriction X_1 = R_1(K_1 \cap K_2) \restriction X_1.$$

So assume that $g \in R_1(K_1) \restriction X_1$. Then $\exists f_1 \in R_1(K_1)$ such that $f_1 \restriction X_1 = g$. We now claim that

$$\forall x \notin X_1 \quad f_1(x) = 1. \tag{4.13}$$

Suppose otherwise and define

$$f'(x) = \begin{cases} f_1(x) & \text{if} \quad x \in X_1 \\ \\ 1 & \text{otherwise.} \end{cases}$$

Then $f' \in K_1$ but $\left| f'^{-1}\{1\} \right| > \left| f_1^{-1}\{1\} \right|$, which is impossible if $f_1 \in R_1(K_1)$. Hence $X_1 \supseteq \{x \mid f_1(x) = 0\}$ (and similarly, $X_2 \supseteq \{x \mid f_2(x) = 0\}$, for $f_2 \in R_1(K_2)$). Thus $\exists f \in K_1 \cap K_2$ such that $f \restriction X_1 = f_1 \restriction X_1$ and $f \restriction X_2 = f_2 \restriction X_2$. Moreover, since $X_1 \cup X_2$ is a support for $K_1 \cap K_2$, can also assume that

$$f(x) = 1, \quad \text{for all} \quad x \notin X_1 \cup X_2. \tag{4.14}$$

Claim now that there is no $h \in K_1 \cap K_2$ such that

$$\left| h^{-1}\{1\} \right| > \left| f^{-1}\{1\} \right|. \tag{4.15}$$

Suppose on the contrary that such an h existed. By (4.14) we may assume $h(x) = 1$ for all $x \notin X_1 \cup X_2$. Notice first that

$$x \in X_1 \cap X_2 \Rightarrow f(x) = h(x). \tag{4.16}$$

To see this, notice that $f \in K_1, h \in K_2$. So $\exists g'$ such that $g' \restriction X_1 = f \restriction X_1$ and $g' \restriction X_2 = h \restriction X_2$. Hence $f(x) = g'(x) = h(x)$, as required. Now,

$$\left|h^{-1}\{1\}\right| = \overbrace{\left|\{y \in X_1 - X_2 \mid h(y) = 1\}\right|}^{\alpha^h} + \left|\{y \in X_2 - X_1 \mid h(y) = 1\}\right| + \\ + \left|\{y \in X_2 \cap X_1 \mid h(y) = 1\}\right|.$$

and

$$\left|f^{-1}\{1\}\right| = \overbrace{\left|\{y \in X_1 - X_2 \mid f(y) = 1\}\right|}^{\alpha^f} + \left|\{y \in X_2 - X_1 \mid f(y) = 1\}\right| + \\ + \left|\{y \in X_2 \cap X_1 \mid f(y) = 1\}\right|.$$

Without loss of generality then, if $|h^{-1}\{1\}| > |f^{-1}\{1\}|$ then $\alpha^h > \alpha^f$. But this leads to the required contradiction. To see that define

$$h'(z) = \begin{cases} h(z) & \text{if} \quad z \in X_1 \\ \\ 1 & \text{otherwise.} \end{cases}$$

Then $h' \in K_1$ but $\left|h'^{-1}\{1\}\right| = |h^{-1}\{1\} \cap X_1| > \left|f_1^{-1}\{1\}\right|$, and this is clearly inconsistent with $f_1 \in R(K_1)$. So $f \in R(K_1 \cap K_2)$ and hence $g \in R(K_1 \cap K_2) \restriction X_1$, as required for this direction of the proof.

As to the other direction for Irrelevance , assume that $g \in R(K_1 \cap K_2) \restriction X_1$ but $g \notin R(K_1) \restriction X_1$. Define

$$g'(x) = \begin{cases} g(x) & \text{if} \quad x \in X_1 \\ \\ 1 & \text{otherwise.} \end{cases}$$

Then, $g' \in K_1$ as it agrees on X_1 with $g \in K_1$. Indeed $g' \notin R(K_1) \restriction X_1$ too, since $g' \restriction X_1 = g \restriction X_1$. Hence $\exists f \in R(K_1)$ such that

$$\left|\{y \in X_1 \mid f(y) = 1\}\right| > \left|\{y \in X_1 \mid g(y) = 1\}\right|. \tag{4.17}$$

Now pick $h \in R(K_1 \cap K_2)$ such that $h \restriction X_1 = g$ and define f' such that $f' \restriction X_1 = f \restriction X_1$ and $f' \restriction X_2 = h \restriction X_2$. As above we can assume that

$$f'(x) = 1 \text{ for all } x \notin X_1 \cup X_2 \tag{4.18}$$

Then $f' \in K_1 \cap K_2$ and $\left|f'^{-1}\{1\} \cap X_1\right| > |h^{-1}\{1\} \cap X_1|$ (by (4.17) and the facts $f' \restriction X_1 = f \restriction X_1$ and $h \restriction X_1 = g$). Thus, since $\left|f'^{-1}\{1\} \cap X_2\right| = |h^{-1}\{1\} \cap X_2|$ and $f' \restriction X_1 \cap X_2 = h \restriction X_1 \cap X_2$, we have that

$$\left|f'^{-1}\{1\} \cap (X_1 \cup X_2)\right| > \left|h^{-1} \restriction \{1\} \cap (X_1 \cup X_2)\right|.$$

But this is inconsistent with the maximality of $|h^{-1}\{1\}|$, concluding the proof of the converse of Theorem 4.2.

■

A pleasing aspect of Theorem 4.2 is that it seems to us to point to precisely the answer(s) that people commonly do come up with when presented with a Rationality-as-conformity choice problem. For example in the case

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

it is our experience that the "fourth" row, 1 1 1 1, is the favored choice. In other words the (unique) choice according to R_1. Of course that is not the only Regulative Reason, R_0 gives $\{0001, 0100, 0010\}$ whilst the trivial reason gives us back the whole set. Clearly though those two Reasons could be seen as inferior to R_1 here because they ultimately require a random choice from a larger set, thus increasing the probability of non-agreement. (This idea will be explored further in the next chapter when we come to Reasons based on Ambiguity.) This seems to point to a further elaboration of our picture whereby the agent might for a particular K experiment with several Reasons and ultimately settle for a choice which depends on K itself. Alternatively one might

hedge one's bets and adopt the "collected extremal choice", $R_{\cup}(K) = R_0(K) \cup R_1(K)$, in the sense of Aizerman and Malishevski (1981) (see also Rott (2001), p. 163) and by the Aizerman-Malishevski Theorem (Theorem 4 of Aizerman and Malishevski (1981)) $R_{\cup}$ is a *Plott function*, that is to say a function that satisfies the so-called Path Independence property introduced in Plott (1973). In the context of the present investigation, however, this route doesn't seem to lead to any interesting development of an alternative Reason.

Of course one might argue about this example that in making the choice of 1 1 1 1 one was not *consciously* aware of any obligation to satisfy Renaming, Obstinacy and Irrelevance. This situation is in fact analogous to the one arising with the conventional theory of rational choice outlined above in section 2.2. There we recalled how the upshot of Savage's framework was that agents whose preferences on acts satisfied the consistency requirements captured by Postulate 1 – Postulate 7 would choose *as if* they where maximising the expected utility of their acts. This analogy in fact runs along two dimensions. Like Savage, we do not claim that the satisfaction of the formal properties constraining the agents' choice corresponds in a strict sense to the actual cognitive processes underlying the act of choice. Yet those constraints correspond to principles which are "justified" on the grounds that, were agents to consciously consider the consequences of transgressing them, they would consider such a transgression inappropriate.

Moreover, it turns out that the general intuitions underlying the common sense principles approach, and hence the Rationality-as-conformity one, are remarkably close to those considered by Carnap (in the context of probabilistic confirmation theory) when developing his programme on Inductive Logic:

> The person X wishes to assign rational credence values to unknown propositions on the basis of the observations he has made. It is the purpose of inductive logic to help him to do that in a rational way; or, more precisely, to give him some general rules, each of which warns him against certain

> unreasonable steps. The rules do not in general lead him to specific values; they leave some freedom of choice within certain limits. What he chooses is not a credence value for a given proposition but rather certain features of a general policy for determining credence values. (Hilpinen, 1973)

Indeed, the principles that characterise commonsense in terms of the Maximum Entropy inference process as well as the principles introduced in this chapter are understood exactly as *policies* helping agents to achieve their goal by forbidding them to undertake certain *unreasonable steps* which, if on the one hand would not lead agents to irrational choices (their degrees of belief are already secured against sure loss by being probabilities) could prevent, on the other hand, agents from making a "more reasonable", i.e. commonsensical, choice.

Thus, be that as it may, we feel that observance of these principles turns out to be both so restrictive and to rather frequently leads to 'the people's choice'. Notice too that if one does adopt a Regulative Reason then one automatically also observes Obstinacy. This *could* then be offered as another defense of Obstinacy against the earlier criticism, that it is no more unreasonable than adopting a Regulative Reason altogether. Whether or not there are alternative sets of "justified" principles which yield interesting families of Reasons such as the one we have considered here remains a matter for further investigation.

We conclude this chapter by showing the already mentioned fact that if we defined a somewhat different choice context, in which we allowed A to be infinite, we would arrive at the following, rather surprising, result.

Proposition 4.12. *If A is infinite and $\mathbb{K}$ is the non-empty subsets of 2^A with finite support, then the trivial reason is the only Regulative Reason.*

Proof. Let R be a Regulative Reason and suppose X is the unique smallest support of $R(2^A)$. Since X is finite and A infinite we can choose a permutation σ such that

$X \cap \sigma^{-1}X = \emptyset$. By Renaming $\sigma^{-1}X$ is a support of $R(2^A)\sigma$ (as noted above) but

$$\begin{aligned} R(2^A)\sigma =& R(2^A\sigma) \quad \text{(by Renaming)} \\ =& R(2^A) \quad \text{(since} \quad 2^A\sigma = 2^A\text{)}. \end{aligned} \tag{4.19}$$

Hence $\sigma^{-1}X$ is also a support of $R(2^A)$, so $X = \sigma^{-1}X$, which is only possible if $X = \emptyset$. Thus $R(2^A) = 2^A$.

Now let $K \in \mathbb{K}$. Then

$$K = K \cap 2^A = K \cap R(2^A) \neq \emptyset$$

so by Obstinacy,

$$R(K) = R(K \cap 2^A) = K \cap R(2^A) = K$$

as required. ■

As we can clearly see from the proof, this result is independent of the principle of Irrelevance.

Chapter 5

The Minimum Ambiguity Reason

ABSTRACT: *Our second characterisation of a choice process facilitating conformity is given in terms of an algorithm for computing the minimally ambiguous (most outstanding) world(s) within a given element of* $\wp^+(2^A)$.

5.1 An informal procedure

In the previous chapter we saw how an agent might arrive at a particular canonical Reason by adopting and adhering to certain principles, principles which (after some consideration) one might suppose any other like-minded agent might similarly come to. An alternative approach, which we shall investigate in this chapter, is to introduce a notion of 'distinguishability', or 'indistinguishability', between elements of K and choose as $R(K)$ those most distinguished, equivalently *least ambiguous*, elements. Instead of being based on principles this $R(K)$ will in the first instance be specified by a procedure, or algorithm, for constructing it.

The idea behind the construction of the Minimum Ambiguity Reason $R_{\mathbb{A}}(K)$ is based on trying to fulfill two requirements. The first requirement is 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 this surely looks like an undesirable situation (and indeed, taking that route may be worse, and will never be better, than avoiding it). So we can sum up the first requirement on $R(K)$ by saying that it should be *closed under the 'indistinguishability relation'*.

The second requirement is that the agent's 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 those desiderata 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 $X_1, X_2, ..., X_k$ at this stage with the same number of elements then the agent could do no worse than combine them into a single potential choice $X_1 \cup X_2 \cup ... \cup X_k$ since the choice of any one of them would be open to the obvious criticism that another 'like-minded agent' could make a different (in this case disjoint) choice, which would not improve the chances of a match (and may make them considerably worse if the first agent subsequently rejected $X_1 \cup X_2 \cup ... \cup X_k$ in favor of a better choice). Faced with this revelation our agent would realize that the 'smallest' way open to reconcile these alternatives is to now permit $X_1 \cup X_2 \cup ... \cup X_k$ as a potential choice whilst dropping $X_1, X_2, ..., X_k$. [Note that this strategy interestingly combines aspects of the "sceptical" as well as the "credulous" approaches to non-monotonic inference recalled above in section 3.1.2.]

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.

In what follows we shall give a formalization of this procedure.

5.2 Permutations and ambiguity

The first step in the construction of the Minimum Ambiguity Reason consists in providing the agent with a notion of equivalence or indistinguishability among worlds in a given choice contest $K \in \wp^+(2^A)$.

In fact with the minimal structure we have available here the notion we want is almost immediate: Elements g, h of K are *indistinguishable* (with respect to K) if there is a permutation σ of A such that

$$K = K\sigma \quad \text{and} \quad g\sigma = h,$$

where as usual $K\sigma = \{ f\sigma \mid f \in K \}$. From now on we shall say that a permutation σ of A *is a permutation of* K if $K = K\sigma$.

The idea here is that within the context of our choice problem a permutation σ of K maps $f \in K$ to an $f\sigma$ in $K\sigma$ which has essentially the standing within $K\sigma$ $(= K)$ as f had within K. In other words as far as K is concerned f and $f\sigma$ are indistinguishable. We shall investigate a more general notion of indistinguishability in section 7.2 below.

The following Lemma is immediate.

Lemma 5.1. *If σ and τ are permutations of K then so are $\sigma\tau$ and σ^{-1}.*

Having now disposed of what we mean by indistinguishability between elements of $K \in \wp^+(2^A)$ we can introduce the key element of this characterisation.

Definition. *For $f \in K$* the ambiguity class of f within K at level m *is recursively defined by:*

$$\mathbb{S}'_0(K, f) = \{ g \in K \mid \exists \text{ permutation } \sigma \text{ of } K \text{ such that } f\sigma = g \},$$

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

This recursive construction captures the idea of 'measuring' the ambiguity of a possible world f within K by, first of all, considering how 'distinguished' (in terms of permutations) f is within K. At each of the subsequent stages we make sure that we do not distinguish among worlds which happen to be in the same ambiguity class. This is fully formalised by introducing the following.

Definition. *For* $f, g \in K$

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

As expected the following can be proved.

Lemma 5.2. $\sim_m$ *is an equivalence relation.*

Proof. By induction on m. For the case $m = 0$ this is clear since if $f, g, h \in K$ and $f\sigma = g$, $g\tau = h$ with σ, τ permutations of K then $g\sigma^{-1} = f$, $f\sigma\tau = h$ and by Lemma 5.1 $\sigma^{-1}, \sigma\tau$ are also permutations of K.

Assume true for m. If $|\mathbb{S}_m(K,f)| > m+1$ then, by the definition of $\mathbb{S}_{m+1}(K,f)$, the result follows immediately from the inductive hypothesis. Otherwise, the reflexivity of $\sim_m$ is again immediate. For symmetry assume that $g \in \mathbb{S}_{m+1}(K,f)$. Then $g \in \{h \in K \mid |\mathbb{S}_m(K,h)| = |\mathbb{S}_m(K,f)|\}$, so $|\mathbb{S}_m(K,g)| = |\mathbb{S}_m(K,f)|$ and $f \in \{h \in K \mid |\mathbb{S}_m(K,h)| = |\mathbb{S}_m(K,g)|\}$. An analogous argument shows that $\sim_{m+1}$ is also transitive. ■

Thus, as f ranges over K, $\sim_m$ induces a partition on K 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 $\mathbb{S}(K,f)$.

The *ambiguity of* f *within* K is then 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:

$$R_{\mathbb{A}}(K) = \{f \in K \mid \forall g \in K,\ \mathbb{A}(K,f) \leq \mathbb{A}(K,g)\}. \tag{5.1}$$

As a rather immediate consequence of the definition of $R_{\mathbb{A}}$ we have the following result.

Proposition 5.3. *$R_{\mathbb{A}}(K) = \mathbb{S}(K,f)$, for any $f \in R_{\mathbb{A}}(K)$*

Proof. Let $f \in R_{\mathbb{A}}(K)$. To show that $\mathbb{S}(K,f) \subseteq R_{\mathbb{A}}(K)$ suppose $\mathbb{S}(K,f) = \mathbb{S}_m(K,f)$ and $g \in \mathbb{S}_m(K,f)$. Then $\mathbb{S}_m(K,g) = \mathbb{S}_m(K,f)$ so $\mathbb{S}_m(K,g)$ must equal $\mathbb{S}(K,g)$ (since m could be taken arbitrarily large) and $|\mathbb{S}(K,g)| = |\mathbb{S}(K,f)|$, so $g \in R_{\mathbb{A}}(K)$. Conversely let $g \in R_{\mathbb{A}}(K)$ and fix some large m. If $g \notin \mathbb{S}(K,f)$, then $\mathbb{S}(K,f) \cap \mathbb{S}(K,g) = \emptyset$ and since both f and g are in $R_{\mathbb{A}}(K)$, then $|\mathbb{S}(K,f)| = |\mathbb{S}(K,g)|$. But this leads to the required contradiction since for m large enough, $|\mathbb{S}_m(K,f)| \leq m+1$ so $\mathbb{S}_m(K,f)$ and $\mathbb{S}_m(K,g)$ would both be proper subsets of $\mathbb{S}_{m+1}(K,f)$. Thus g would eventually be in $\mathbb{S}_m(K,f)$, contradicting the hypothesis. ■

Example 9. Let $K \in \mathbb{K}$ and suppose that as f ranges over K the 0-ambiguity classes of f in K are given by the following partition of K

$\{a_1, a_2\}, \{b_1, b_2\}, \{c_1, c_2\},$

$\{d_1, d_2, d_3\}, \{e_1, e_2, e_3\},$

$\{f_1, f_2, ..., f_6\}, \{g_1, g_2, ..., g_6\},$

$\{h_1, h_2, ..., h_{12}\},$

$\{i_1, i_2, ..., i_{24}\}.$

For $m = 1$ the classes remain fixed. For $m = 2$ the first three classes get combined and the $\mathbb{S}_2(K,f)$ look like

$\{a_1, a_2, b_1, b_2, c_1, c_2\},$

$\{d_1, d_2, d_3\}, \{e_1, e_2, e_3\},$

$\{f_1, f_2, ..., f_6\}, \{g_1, g_2, ..., g_6\},$

$\{h_1, h_2, ..., h_{12}\},$

$\{i_1, i_2, ..., i_{24}\}.$

Similarly for $m = 3$ where the two classes of size 3 are combined so that the $\mathbb{S}_3(K, f)$ become

$\{a_1, a_2, b_1, b_2, c_1, c_2\}$,

$\{d_1, d_2, d_3, e_1, e_2, e_3\}$,

$\{f_1, f_2, ..., f_6\}, \{g_1, g_2, ..., g_6\}$,

$\{h_1, h_2, ..., h_{12}\}$,

$\{i_1, i_2, ..., i_{24}\}$.

The ambiguity classes do not change until step 6 when the four classes with 6 elements are combined making $\mathbb{S}_6(K, f)$ look like

$\{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, d_3, e_1, e_2, e_3, f_1, f_2, ..., f_6, g_1, g_2, ..., g_6\}$,

$\{h_1, h_2, ..., h_{12}\}$,

$\{i_1, i_2, ..., i_{24}\}$.

Finally, we combine the two classes with 24 elements and obtain $\mathbb{S}_{24}(K, f)$ with just two classes

$\{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, d_3, e_1, e_2, e_3, f_1, f_2, ..., f_6, g_1, g_2, ..., g_6, i_1, i_2, ..., i_{24}\}$,

$\{h_1, h_2, ..., h_{12}\}$.

Clearly the ambiguity classes stabilize at this 24-th step and hence the Minimum Ambiguity Reason for this K gives the 12-set $\{h_1, h_2, ..., h_{12}\}$.

Notice that, in the definition of the ambiguity classes of K, the splitting of the inductive step into two cases is indeed necessary to ensure that some sets closed under permutations of K are not dismissed unnecessarily early. This same example shows that if we allowed the inductive step in the definition to be replaced by the (somewhat more intuitive) equation

$$\mathbb{S}_{m+1}(K, f) = \{g \in K \mid |\mathbb{S}_m(K, f)| = |\mathbb{S}_m(K, g)|\} \tag{5.2}$$

we would fail to pick the "obvious" smallest such subset of K. To see this suppose again that K is as above but this time the alternative procedure based on (5.2) was used to construct $R_{\mathbb{A}}$. Then we would have all the classes of the same size merged in one step so that the $1-$ambiguity classes $\mathbb{S}_1(K, f)$ would look like:

$\{a_1, a_2, b_1, b_2, c_1, c_2\}$,

$\{d_1, d_2, d_3, e_1, e_2, e_3\}$,

$\{f_1, f_2, ..., f_6, g_1, g_2, ..., g_6\}$,

$\{h_1, h_2, ..., h_{12}\}$,

$\{i_1, i_2, ..., i_{24}\}$.

Then $\mathbb{S}_2(K, f)$ would look like this:

$\{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, d_3, e_1, e_2, e_3\}$,

$\{f_1, f_2, ..., f_6, g_1, g_2, ..., g_6, h_1, h_2, ..., h_{12}\}$,

$\{i_1, i_2, ..., i_{24}\}$,

so that the procedure stabilizes at $m = 3$ with $\mathbb{S}(K, f)$ of the form:

$\{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, d_3, e_1, e_2, e_3\}$,

$\{f_1, f_2, ..., f_6, g_1, g_2, ..., g_6, h_1, h_2, ..., h_{12}, i_1, i_2, ..., i_{24}\}$,

Hence, the construction that follows the alternative definition of ambiguity classes, which imposes no restriction on appropriate stage for the combination of the classes, leads again to a 12-set. However, this alternative procedure appears to miss out what naturally seems to be a more distinguished subset of K.

5.3 Justifying the Minimum Ambiguity Reason

We now want to show that the Minimum Ambiguity Reason defined in (5.1) is an adequate formalization of the informal description given in section 5.1. Recall that we put forward two informal desiderata for the resulting selection from K, firstly that it should be closed under indistinguishability and secondly that it should be the unique smallest possible such subset not eliminated by there being a like-minded agent who by similar reasoning could arrive at a different answer.

As far as the former is concerned notice that by proposition 5.3 $R_{\mathbb{A}}(K)$ is closed under all the $\sim_m$, not just $\sim_0$. Thus this requirement of closure under indistinguishability is met, *assuming* of course that one accepts this interpretation of 'indistinguishability'. Indeed $R_{\mathbb{A}}$ satisfies Renaming as we now show.

Theorem 5.4. *$R_{\mathbb{A}}$ satisfies Renaming.*

Proof. As usual let σ be a permutation of A. We need to prove that

$$R_{\mathbb{A}}(K)\sigma = R_{\mathbb{A}}(K\sigma).$$

We first show by induction on m that for all $f \in K$, $\mathbb{S}_m(K,f)\sigma = \mathbb{S}_m(K\sigma, f\sigma)$. To show the base case $m = 0$ for all $f \in K$, let

$$\mathbb{S}_0(K,f) = \{\, g_1, \ldots, g_q \,\}.$$

Choose a permutation τ of K such that $f\tau = g_i$. Then $\sigma^{-1}\tau\sigma$ is a permutation of $K\sigma$ and $(f\sigma)\sigma^{-1}\tau\sigma = g_i\sigma$. Hence, $\mathbb{S}_0(K,f)\sigma \subseteq \mathbb{S}_0(K\sigma, f\sigma)$. Similarly, $\mathbb{S}_0(K\sigma, f\sigma)\sigma^{-1} \subseteq \mathbb{S}_0(K,f)$, so equality must hold here.

Assume now the result for the $\mathbb{S}_m$-th ambiguity class, so we want to prove that

$$\mathbb{S}_{m+1}(K,f)\sigma = \mathbb{S}_{m+1}(K\sigma, f\sigma).$$

We distinguish between two cases, corresponding to the ones appearing in the construction of the ambiguity classes. Recall that $\mathbb{S}_{m+1}(K,f) = \mathbb{S}_m(K,f)$ if $m+1 > |\mathbb{S}_m(K,f)|$. So, in this case, the result follows immediately by the inductive hypothesis. Otherwise, since σ (on 2^A) is 1-1, it is enough to see that

$$\begin{aligned}
\mathbb{S}_{m+1}(K,f)\sigma &= \{g \in K \mid |\mathbb{S}_m(K,f)| = |\mathbb{S}_m(K,g)|\}\sigma \\
&= \{g\sigma \in K\sigma \mid |\mathbb{S}_m(K\sigma, f\sigma)| = |\mathbb{S}_m(K\sigma, f\sigma)|\} \ \text{(i.h.)} \\
&= \mathbb{S}_{m+1}(K\sigma, f\sigma).
\end{aligned}$$

Since, by Lemma 5.3, $R_{\mathbb{A}}(K)$ is the smallest $\mathbb{S}(K,f)$, this concludes the proof of the Lemma. ■

Before further considering how far our formal construction of $R_{\mathbb{A}}(K)$ matches the informal description in section 3.1, it will be useful to have the next result to hand.

Theorem 5.5. *A non-empty $K' \subseteq K$ is closed under permutations of K into itself if and only if there exists a Reason R satisfying Renaming such that $R(K) = K'$.*

Proof. The direction from right to left follows immediately from the Renaming principle. For the other direction define, for $K_1 \subseteq 2^A, K_1 \neq \emptyset$,

$$R(K_1) = \begin{cases} K'\sigma & \text{if } K_1 = K\sigma \text{ for some permutation } \sigma \text{ of } A; \\ K_1 & \text{otherwise .} \end{cases} \tag{5.3}$$

Note that in the first case $R(K_1)$ is defined unambiguously, that is to say, whenever we have two permutations σ_1, σ_2 of A such that $K_1 = K\sigma_1 = K\sigma_2$, then $K'\sigma_1 = K'\sigma_2$. This follows since in this case, $\sigma_2\sigma_1^{-1}$ is a permutation of A and $K\sigma_2\sigma_1^{-1} = K$ so $K'\sigma_2\sigma_1^{-1} = K'$, i.e. $K'\sigma_1 = K'\sigma_2$.

We now want to show that if σ is a permutation of A and $K_1\sigma = K_2$ then $R(K_2) = R(K_1)\sigma$. If K_1 is covered by the first case of (5.3), then so is K_2, for if τ is a permutation of A such that $K_1 = K\tau$, then $K_2 = K\tau\sigma$ and $R(K_1\sigma) = R(K_2) = K'\tau\sigma = R(K_1)\sigma$. If K_1 is covered by the second case of (5.3), so is K_2 since if $K_2 = K\tau$ for some permutation τ of A, then $K_1 = K\tau\sigma^{-1}$ so $R(K_1)$ would be defined by the first case. It follows then that here we must have $R(K_1\sigma) = R(K_2) = K_2 = K_1\sigma = R(K_1)\sigma$ as required. ■

The importance of this result 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 Renaming, 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 Renaming. 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 of section 3.1.

5.4 Comparing Regulative and Minimum Ambiguity Reasons

In this and the previous chapter we have put forward arguments for both the Regulative and Minimum Ambiguity Reasons being considered as "rational" choice processes with respect to the conformity problem. It is interesting to note, however that in practice neither seems to come out self evidently better in all cases. For example, in the case considered earlier of

$$\begin{matrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{matrix}$$

R_1 gives the singleton $\{1111\}$ whilst $R_{\mathbb{A}}$ gives the somewhat unexceptional $\{0011, 1100\}$ and R_0 the rather useless $\{0011, 0110, 1100\}$. On the other hand if we take the subset

$$\begin{matrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{matrix}$$

of this set $R_{\mathbb{A}}$ gives $\{0110\}$ whilst both R_1 and R_0 give the whole set. And in fact we noted earlier that one might argue against Obstinacy by saying that after applying a Regulative Reason to K, 0011 could be seen to gain a somehow distinguished status and hence should be selected violating Obstinacy. We can now refine our rejection of this argument by saying that this violation of Obstinacy would require an agent to change Reason "on the fly", namely passing from the Regulative to the Minimum Ambiguity Reason *within the same choice problem*. The argument for rejecting this possibility is that if this were allowed, sufficiently large K's would generate an explosion of possible combinations of Reasons which in all probability would put back the agents in a position of choosing randomly from K!

Concerning the defining principles of the Regulative Reasons, whilst as we have seen $R_{\mathbb{A}}$ does satisfy Renaming, the above example shows that it fails to satisfy

Obstinacy. Indeed a simple example shows that it does not even satisfy Idempotence, that is $R(R(K)) = R(K)$, an immediate consequence of Obstinacy.

Example 10. Let K be

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

Here $R_{\mathbb{A}}(K)$ gives $\{110000, 100000, 010000\}$ whereas $R_{\mathbb{A}}(R_{\mathbb{A}}(K))$ returns $\{110000\}$. It is interesting to note that, as far as the interest in "small size" goes, this failure of Idempotence can be indeed welcome. In fact agents might exploit this fact and keep applying $R_{\mathbb{A}}$ until the set $R_{\mathbb{A}}(K)$ stabilises. A consequence of this being that agents will end up to randomising (when Idempotence fails) over a *smaller* set of possible worlds, thus increasing their chances of conforming.

Finally $R_{\mathbb{A}}$ does not satisfy Irrelevance either. For an example to show this let K_1 consist of

$$\begin{array}{cccccccccccc} 1 & 0 & 0 & 1 & * & * & * & * & * & * & * \\ 1 & 1 & 0 & 1 & * & * & * & * & * & * & * \\ 1 & 1 & 1 & 1 & * & * & * & * & * & * & * \\ 1 & 0 & 0 & 0 & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 1 & * & * & * & * & * & * & * \end{array}$$

and let K_2 consist of

$$
\begin{array}{cccccccccccc}
* & * & * & * & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
* & * & * & * & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
* & * & * & * & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
* & * & * & * & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
* & * & * & * & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
* & * & * & * & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
* & * & * & * & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}
$$

where $*$ indicates a free choice of 0 or 1. Then K_1, K_2 satisfy the requirements of Irrelevance and $R_{\mathbb{A}}(K_1), R_{\mathbb{A}}(K_2)$ are respectively

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

whereas $R_{\mathbb{A}}(K_1 \cap K_2)$ is

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

Chapter 6

The Smallest Uniquely Definable Reason

ABSTRACT: *The third approach to the characterisation of reasons facilitating conformity is model-theoretic. Given an adequate structure, commonsensical agents should choose the smallest (first-order) uniquely definable subset of the initial set of possible worlds.*

6.1 The model theoretic structure of the main problem

In this chapter we present another Reason which, at first sight, looks a serious challenger to the Regulative and Minimum Ambiguity Reasons so far introduced.

Consider again the main problem of Rationality-as-conformity, i.e. an agent who is given a non-empty subset K of 2^A from which to attempt to make a choice which is common to another like-minded yet inaccessible agent. A natural approach here might be for the agent to consider all non-empty subsets of K that could be described, or to use a more formal term, defined, within the *structure available* to the agent. If some individual element was definable (meaning definable in this structure *without parameters*) then this would surely be a natural choice, unless of course there were

other such elements. Similarly choosing a small definable set and then choosing randomly from within it would seem a good strategy, *provided there were no other definable sets of the same size.* Reasoning along these lines then suggests that our agent could reach the conclusion that s/he should choose the smallest definable set for which there was no other definable set of the same size.

Of course all this depends on what we take to be the *structure available* to the agent. In what follows we shall consider the case when the agent can recognize 0 and 1, elements of A, $\{0,1\}$ and K, composition and equality. Precisely, let $\mathcal{M}$ be the structure

$$\langle \{0,1\} \cup A \cup K, \{0,1\}, A, K, =, Comp, 0, 1 \rangle$$

where $=$ is equality for $\{0,1\} \cup A \cup K$ (we assume of course that A, $\{0,1\}$, 2^A are all disjoint) and $Comp$ is a binary function which on $f \in K$, $a \in A$ gives $f(a)$ (and, say, the first coordinate on arguments not of this form). As usual we shall write $f(a) = i$ in place of $Comp(f, a) = i$ etc..

Note that at this point one might argue that the agent could then also recognize automorphisms of $\mathcal{M}$ so the set of these too should be added to our structure, and the whole process repeated, and repeated .... In fact this does not change the definable subsets of K so it turns out there is no point in going down this path.

6.2 The Uniquely Smallest Definable Reason characterised

We define the Uniquely Smallest Definable Reason, $R_{\mathbb{U}}$, by setting $R_{\mathbb{U}}(K)$ to be that smallest $\emptyset \neq K' \subseteq K$ first order definable in $\mathcal{M}$ for which there is no other definable subset of the same size.

The results that follow are directed towards understanding the structure of $R_{\mathbb{U}}(K)$ and its relationship to $R_{\mathbb{A}}(K)$.

Lemma 6.1. *Every permutation σ_0 of K determines an automorphism j_{σ_0} of $\mathcal{M}$ given by the identity on $\{0,1\}$ and*

$$a \in A \longmapsto \sigma_0^{-1}(a), \tag{6.1}$$

and

$$f \in K \longmapsto f\sigma_0. \tag{6.2}$$

Conversely every automorphism j_0 of $\mathcal{M}$ determines a permutation σ_{j_0} of K given by

$$\sigma_{j_0}(a) = j_0^{-1}(a) \tag{6.3}$$

for $a \in A$.

Furthermore for $f \in K$, $f\sigma_{j_0} = j_0(f)$ and the corresponding automorphism determined by $j_{\sigma_{j_0}}$ is j_0 again.

Proof. For σ_0 a permutation of K it is clear that j_{σ_0} defined by (6.1) and (6.2) gives a 1-1 onto mapping from A and K into themselves. All that remains to show this first part is to notice that by direct substitution,

$$j_{\sigma_0}(Comp(f,a)) = Comp(f,a) = f(a) = f\sigma_0(\sigma_0^{-1}(a)) = Comp(j_{\sigma_0}(f), j_{\sigma_0}(a)).$$

In the other direction let j_0 be an automorphism of $\mathcal{M}$ and define σ_{j_0} by (6.3). Then since j_0 is an automorphism of $\mathcal{M}$, σ_{j_0} is a permutation of A and for $f \in K$, $a \in A$,

$$f(a) = j_0(f(a)) = j_0(Comp(f,a)) = Comp(j_0(f), j_0(a)),$$

equivalently,

$$f(a) = j_0(f)(j_0(a)) = j_0(f\sigma_{j_0}^{-1})(a).$$

Hence

$$j_0^{-1}(f)(a) = f\sigma_{j_0}^{-1}(a)$$

so $\sigma_{j_0}^{-1}$ (and hence σ_{j_0} by Lemma 5.1) is a permutation of K since $j_0^{-1}(f) \in K$, as required.

The last part now follows immediately from the definitions (6.1), (6.2), (6.3). ■

We say that $K' \subseteq K$ satisfies Renaming within K if for all permutations σ of K, $K' = K'\sigma$. Thus 'standard Renaming' is just Renaming within 2^A.

Theorem 6.2. *A non-empty subset K' of K is definable (without parameters) in $\mathcal{M}$ if and only if K' satisfies Renaming within K.*

Proof. Suppose that K' is definable in $\mathcal{M}$. Then clearly K' is fixed under all automorphisms of $\mathcal{M}$. In particular if σ is a permutation of K then by Lemma 6.1 j_σ is an automorphism of $\mathcal{M}$ so

$$K' = j_\sigma(K') = K'\sigma$$

Conversely suppose that K' satisfies Renaming within K. Then since every automorphism of $\mathcal{M}$ is of the form j_σ for some permutation σ of K and $j_\sigma(K') = K'\sigma = K'$ it follows that K' is fixed under all automorphisms of $\mathcal{M}$. Consider now the types $\theta_1^i(x), \theta_2^i(x), \theta_3^i(x), ...$ of the elements f_i of K in $\mathcal{M}$. If there were $f_i \in K'$ and $f_j \notin K'$ with the same type then by a back and forth argument (see e.g. Marker, 2002) we could construct an automorphism of $\mathcal{M}$ sending f_i to f_j, contradicting the fact that K' is fixed under automorphisms. It follows that for some n the formulae

$$\theta_1^i(x) \wedge \theta_2^i(x) \wedge ... \wedge \theta_n^i(x)$$

and

$$\theta_1^j(x) \wedge \theta_2^j(x) \wedge ... \wedge \theta_n^j(x)$$

are mutually contradictory when $f_i \in K'$ and $f_j \notin K'$. From this it clearly follows that the formula

$$\bigvee_{f_i \in K'} \bigwedge_{m=1}^{n} \theta_m^i(x)$$

defines K' in $\mathcal{M}$ for suitably large n. ■

Corollary 6.3. *The sets $\mathbb{S}_m(K, f)$ are definable in $\mathcal{M}$*

Proof. These sets are clearly closed under permutations of K so the result follows from Theorem 6.2. ■

Theorem 6.4. *For all $K \in \mathbb{K}$, $|R_{\mathbb{A}}(K)| \leq |R_{\mathbb{U}}(K)|$, with equality just if $R_{\mathbb{A}}(K) = R_{\mathbb{U}}(K)$.*

Proof. We shall show that for all m. If $f \in R_{\mathbb{U}}(K)$ then $\mathbb{S}_m(K, f) \subseteq R_{\mathbb{U}}(K)$. For $m = 0$ this is clear since $R_{\mathbb{U}}(K)$, being definable must be closed under permutations of K. Assume the result for m and let $f \in R_{\mathbb{U}}(K)$. If $\mathbb{S}_{m+1}(K, f)$ were not a subset of $R_{\mathbb{U}}(K)$ there would be $g \in K$ such that $|\mathbb{S}_m(K, f)| = |\mathbb{S}_m(K, g)|$ but $g \notin R_{\mathbb{U}}(K)$. Indeed $\mathbb{S}_m(K, g)$ would have to be entirely disjoint from $R_{\mathbb{U}}(K)$ by the inductive hypothesis. By Corollary 6.3 $\mathbb{S}_m(K, f)$ and $\mathbb{S}_m(K, g)$ are both definable, and hence so is

$$R_{\mathbb{U}}(K) \cup \mathbb{S}_m(K, g) - \mathbb{S}_m(K, f).$$

But this set is different from $R_{\mathbb{U}}(K)$ yet has the same size, contradiction.

Having established this fact we notice that for $f \in R_{\mathbb{U}}(K)$ we must have $\mathbb{S}(K, f) \subseteq R_{\mathbb{U}}(K)$ so since $R_{\mathbb{A}}(K)$ is the smallest of the $\mathbb{S}(K, g)$ the result follows. ■

In a way Theorem 6.4 is rather surprising in that one might initially have imagined that $R_{\mathbb{U}}(K)$, by its very definition, was about as specific a set as one could hope to describe. That $R_{\mathbb{A}}(K)$ can be strictly smaller than $R_{\mathbb{U}}(K)$ can be seen from the case when the $\sim_0$ equivalence classes look like

$$\{a_1, a_2\}, \{b_1, b_2\}, \{c_1, c_2\}, \{d_1, d_2, d_3, d_4\}.$$

In this case $R_{\mathbb{A}}$ gives $\{d_1, d_2, d_3, d_4\}$ whereas $R_{\mathbb{U}}$ just gives the union of all these sets as each pairwise union of the above classes is in fact definable.

We now briefly consider the relationship between the Regulative Reasons and $R_{\mathbb{U}}$. Since the set

$$R_i(K) = \{\, f \in K \mid \forall\, g \in K, |f^{-1}(i)| \geq |g^{-1}(i)| \,\}.$$

is definable in $\mathcal{M}$ $R_i(K)$ *is* a candidate for $R_{\mathbb{U}}(K)$. So if $|R_i(K)| < |R_{\mathbb{U}}(K)|$ it must be the case that there is another definable subset of $\mathcal{M}$ with the same size as $R_i(K)$. If $|R_i(K)| = |R_{\mathbb{U}}(K)|$ then in fact $R_i(K) = R_{\mathbb{U}}(K)$. From this point of view then $R_{\mathbb{U}}$ (and by Theorem 6.4 also $R_{\mathbb{A}}$) might be seen to be always at least as satisfactory as the

R_i. On the other hand the R_i are in a practical sense computationally undemanding. [The computational complexity of the relation $f \sim_0 g$ between elements of K is currently unresolved, which strongly suggests that even if a polynomial time algorithm does exist it is far from transparent.]

We finally remark that, using the same examples as for $R_{\mathbb{A}}$, $R_{\mathbb{U}}$ also fails Obstinacy and Irrelevance.

Chapter 7

Variations on the theme

Abstract: *We investigate an analogue of the Regulative Reasons in the case of probabilistic possible worlds. Then a generalisation of the Minimum Ambiguity construction is discussed.*

The purpose of this chapter is two-fold. In the first part we shall consider a more sophisticated notion of possible worlds than the one adopted so far in the formalisation of Rationality-as-conformity. In this revised formal setting we shall be able to compare a probabilistic analogue of the Regulative Reasons to the Paris-Vencovská characterization. As a consequence of the main result of this part, theorem 7.2, this initial comparison fails to be encouraging.

In the second part of the chapter we shall consider a generalisation of the Minimum Ambiguity construction introduced above. This will be based on a notion of indistinguishability according to which the structure of the choice problem is invariant under permutations of the set $\{0, 1\}$.

7.1 Probabilistic possible worlds

Much of our motivation for the investigation of Rationality-as-conformity came from the Paris-Vencovská characterisation of common sense in probabilistic logic and the desire to explain *why* it was that its underlying principles where considered 'common sense'. The Regulative Reasons characterised above could be said to supply such

an explanation in a particularly simple situation. Certainly the resulting answers seem to agree with rationality or common sense as conformity, though perhaps the informal justifications given for Obstinacy and Irrelevance could be more convincing. Nevertheless it would certainly be pleasing to press on and extend this explanation to common sense as formulated for probabilistic uncertain reasoning.

Our initial investigations in this direction suggest that there are a number of difficulties in extending the Regulative approach to the probabilistic case. In order to illustrate some of these points, we need to fix a suitable framework to represent our Rationality-as-conformity main problem. A somehow natural choice would be to take probability functions on the sentences of a finite propositional language, as in the Paris-Vencovská framework, as possible worlds. However, this route would not allow us to have the property corresponding to (finite) support, necessary for the characterisation of the Regulative Reasons. In order to cater for this property one could consider the equivalent framework of probability functions defined on the Lindenbaum algebra of the atoms of L and hence, by further abstraction, by defining probabilistic possible worlds on a finite Boolean algebra $\mathbb{B}$.

More precisely, we shall say, as usual, that a map w on $\mathbb{B}$ is a probability function

$$w : \mathbb{B} \longrightarrow [0,1]$$

if the following are satisfied:

$$w(\mathbf{1}) = 1 \quad \text{and} \tag{7.1}$$

$$\text{If } b_1 \wedge b_2 = \mathbf{0} \text{ then } w(b_1 \vee b_2) = w(b_1) + w(b_2), \tag{7.2}$$

where $b_1, b_2 \in \mathbb{B}$ and $\mathbf{1}$ and $\mathbf{0}$ are the top and bottom elements of $\mathbb{B}$, respectively.

In analogy with the Paris-Vencovská characterisation then, the set of probabilistic possible worlds $\mathbb{P}$ can be taken to be the set of all probability functions w such that

$$\sum_{\alpha_i} w(\alpha_i) = 1,$$

where the α_i run through the (finitely many) atoms of $\mathbb{B}$. Consideration of the sort of knowledge bases investigated there, where knowledge is represented in terms of

linear constraints on a probability function (see e.g. Paris, 1994, ch.2), suggests that the notion of a support for K in $\mathbb{K}$ might now be replaced by *subalgebra support*.

For $\mathbb{B}$ a finite Boolean algebra we say as usual that B is a subalgebra of $\mathbb{B}$ if $\mathbf{0} \in B$ and B is closed under meet and complement. In this case we can go on and define the choice contexts of the Rationality-as-conformity situation analogously to the case studied before, that is we identify the elements K of the set $\mathbb{C}$ with the non-empty subsets of $\mathbb{P}$ of the form

$$\left\{ w \in \mathbb{P} \mid \sum_{\alpha \in \mathbb{B}} a_{i\alpha} w(\alpha) = b_i, i = 1, 2, \ldots, r \right\},$$

where $a_{i\alpha}, b_i \in \mathbb{R}$.

Definition. *For $K \in \mathbb{C}$ we say that the* subalgebra B of $\mathbb{B}$ is a support of K *if for all $w_1, w_2 \in \mathbb{P}$, if $w_1 \restriction B = w_2 \restriction B$ then*

$$w_1 \in K \iff w_2 \in K.$$

As the following (presumably known) result shows, each such K has a unique smallest such subalgebra support.

Proposition 7.1. *If $K \in \mathbb{C}$ then K has a smallest support.*

Proof. Let B and C be subalgebras of $\mathbb{B}$, with atoms $b_1, b_2, \ldots, b_k$ and $c_1, c_2, \ldots, c_m$ respectively, and suppose that B and C are both subalgebra supports for K. The claim is that $B \cap C$ is a subalgebra support for K. So we assume that f_1, f_2 agree on $B \cap C$ and want to show that

$$f_1 \in K \iff f_2 \in K.$$

Let $d_1, d_2, \ldots, d_j$ be the atoms of $D = B \cap C$ and for $\alpha \leq d_i$ an atom of $\mathbb{B}$ define

$$f(\alpha) = f_1(d_i)/|d_i|$$

where $|d_i|$ is the number of atoms of $\mathbb{B}$ contained in d_i. Notice we get the same function if we replace f_1 here by f_2.

To prove the result it is enough to show the existence of $g_1, g_2, \ldots, g_r$ such that $g_1 = f_1$, $g_r = f$ and for each $i = 1, 2, \ldots, r-1$ g_i agrees with g_{i+1} either on B or on C (so certainly they all agree on D). [This is enough because it forces

$$f_1 \in K \iff g_2 \in K \iff g_3 \in K \iff \ldots \iff g_{r-1} \in K \iff f \in K$$

and similarly starting with f_2.] The idea is to produce such $g_1, g_2, \ldots$ so that at each stage

$$|\{\alpha \mid g_i(\alpha) = f(\alpha)\}| < |\{\alpha \mid g_{i+1}(\alpha) = f(\alpha)\}|. \tag{7.3}$$

Clearly this will work if at each stage i for which $g_i \neq f$ we can find such a g_{i+1}. So suppose we have such a g_i and without loss of generality $g_i(\alpha) < f(\alpha)$ with $\alpha \leq d_n$. Then there must be an atom $\beta \leq d_n$ such that $g_i(\beta) > f(\beta)$, since for γ ranging over the atoms of the overlying algebra $\mathbb{B}$,

$$\sum_{\gamma \leq d_n} g_i(\gamma) = g_i(d_n) = f_1(d_n) = f(d_n) = \sum_{\gamma \leq d_n} f(\gamma).$$

Since α and β are both dominated by the same d_n there must be some $e_1, e_2, e_3, e_4, \ldots, e_k$, where the e_m are alternately atoms of B and C, and atoms of $\mathbb{B}$ $\alpha_i \leq e_i \cap e_{i+1}$ with $\alpha = \alpha_1 \leq e_1$ and $\beta \leq e_k$. To see that such $e_1, e_2, e_3, e_4, \ldots, e_k$ must exist, let

$$m = \bigvee \{\beta \mid \exists e_1, e_2, e_3, e_4, \ldots, e_k \text{ as above}\}.$$

Then the set m is exactly the element of $B \cap C$ containing α. In fact $d_n \subseteq m$ since $\emptyset \neq \alpha \subseteq m \cap d_n$ and $m \in B$, $m \in C$. To see that this follows suppose on the contrary that $m \notin B$. Then there would be an atom e of B with $\emptyset \neq e \cap m < e$ while, by the construction of m, $e \cap m = e$. (Similarly for $m \in C$.) The converse inclusion $m \subseteq d_n$ is also true since $d_n \in B \cap C$ and therefore atoms e_m will satisfy that if $\emptyset \neq e_m \cap d_n$ then $e_m \leq d_n$, so they will all be in d_n.

Then somewhere along this path

$$(\alpha =)\alpha_1, \alpha_2, \ldots, \alpha_{k-1}, \beta$$

there must be a consecutive pair γ, δ with $g_i(\gamma) > f(\gamma) = f(\delta) > g_i(\delta)$. Pick such a pair and define g_{i+1} to agree everywhere with g_i except that

$$g_{i+1}(\gamma) = f(\gamma),\ g_{i+1}(\delta) = g_i(\delta) + g_i(\gamma) - f(\gamma).$$

Then since γ, δ are both dominated by of one of the e_g and e_g is either an atom of B or an atom of C, g_{i+1} either agrees with g_i on B or on C (as required) and clearly (7.3) holds, as required.

■

Within this framework we can now press ahead and define the Regulative principles of probabilistic possible worlds. By doing so we can compare directly the resulting characterisation with Paris-Vencovská one.

Renaming:

Let $K \in \mathbb{C}$ and let j be an automorphism of $\mathbb{B}$ (i.e. permutations of the atoms). R satisfies *Renaming* if whenever $Kj = \{wj \,|\, w \in K\}$ then

$$R(Kj) = R(K)j.$$

Obstinacy:

Let $K_1, K_2 \in \mathbb{C}$. R satisfies *Obstinacy* if whenever $R(K_1) \cap K_2 \neq \emptyset$ then

$$R(K_1 \cap K_2) = R(K_1) \cap K_2.$$

Irrelevance:

Suppose $K_1, K_2 \in \mathbb{C}$ with supports B_1, B_2 respectively and for any $w_1 \in K_1$ and $w_2 \in K_2$ there exists $w_3 \in \mathbb{P}$ such that $w_3 \restriction B_1 = w_1 \restriction B_1$ and $w_3 \restriction B_2 = w_2 \restriction B_2$. Then

$$R(K_1) \restriction B_1 = R(K_1 \cap K_2) \restriction B_1$$

where

$$R(K) \restriction B = \{\, w \restriction B \,|\, w \in R(K)\}.$$

The trivial Reason surely satisfies the above principles so the question now is to investigate the class of Reasons characterised by this probabilistic version of the Regulative principles in relation to common sense *à la* Paris-Vencovská . Unfortunately, however, the comparison fails to be encouraging. Recall (Theorem 1.1 of section 1.2 above) that there is a unique inference process satisfying those common sense principles given in Paris and Vencovská (1990, 2001), namely the *Maximum Entropy* inference Process, ME. It turns out, however, that ME *does not* satisfy the version of Irrelevance defined above. Indeed a somewhat stronger result can be proved that probabilistic Regulative Reasons cannot be singleton valued on certain knowledge bases, a property which ME does happen to satisfy.

Theorem 7.2. *Assume that* $\mathbb{B}$ *has at least 8 atoms and that* R *satisfies the probabilistic version of Renaming, Obstinacy and Irrelevance. Then provided* $\mathbb{C}$ *contains certain non-empty subsets* K *of* $\mathbb{P}$ *(specified in the proof) there are some such* K *for which* $|R(K)| > 1$.

Proof. Suppose on the contrary that this does not hold. Let $b_1, b_2, ..., b_6$ be disjoint non-zero elements of $\mathbb{B}$ with sup the top element of $\mathbb{B}$ such that b_2, b_3, b_4, b_5 all have the same number of atoms less than or equal to them whilst b_1, b_6 each have twice that number. [This is possible by the assumption on $\mathbb{B}$ and the fact that in any finite Boolean Algebra the number of atoms is a power of 2.] Let K_1 be the set

$$\{w \in \mathbb{P} \,|\, w(b_1) = 1/16,\, w(b_2 \vee b_3) = 9/16, w(b_4 \vee b_5 \vee b_6) = 3/8\,\}.$$

Using the assumption let $R(K_1) = \{w_1\}$. By Renaming and the assumption on $b_1, ..., b_6$ it follows that we must have

$$w_1(b_2) = w_1(b_3) = 9/32,$$

$$w_1(b_4) = w_1(b_5) = w_1(b_6)/2 = 3/32.$$

Now let K_2 be the set

$$\{w \in \mathbb{P} \,|\, w(b_6) = 3/8 - x,\, w(b_3 \vee b_5) = 2x, w(b_1 \vee b_2 \vee b_4) = 5/8 - x,\ x \in [0, 3/8]\,\}.$$

Notice that $w_1 \in K_2$ (take $x = 3/16$). Hence by Obstinacy $R(K_1 \cap K_2) = \{w_1\}$.

Let $R(\mathbb{P}) = \{w_2\}$. Then

$$w_2(b_1) = 1/4,\ w(b_2 \vee b_3) = 1/4, w(b_4 \vee b_5 \vee b_6) = 1/2$$

so $w_2 \in K_2$. By Obstinacy then, $R(K_2) = \{w_2\}$.

But now it happens that K_1 and K_2 satisfy the requirements of Irrelevance since if $w' \in K_1, w'' \in K_2$, say for a particular $x \in [0, 3/8]$, then w agrees with w' on $b_1, b_2 \vee b_3$ and $b_4 \vee b_5 \vee b_6$ and agrees with w'' on $b_6, b_3 \vee b_5$, $b_1 \vee b_2 \vee b_4$ where w is defined as follows:

if $2x \leq 9/16$ then

$$w(b_6) = 3/8 - x,\ w(b_5) = 0,\ w(b_3) = 2x,$$

$$w(b_4) = x,\ w(b_2) = 9/16 - 2x,\ w(b_1) = 1/16,$$

whilst if $2x \geq 9/16$ then

$$w(b_6) = 3/8 - x,\ w(b_5) = 2x - 9/16,\ w(b_3) = 9/16,$$

$$w(b_4) = 9/16 - x,\ w(b_2) = 0,\ w(b_1) = 1/16.$$

By Irrelevance, and our assumption, it follows that since $R(K_1 \cap K_2) = \{w_1\}$ and $R(K_2) = \{w_2\}$, w_1 must agree with w_2 on the subalgebra generated by $b_6, b_3 \vee b_5, b_1 \vee b_2 \vee b_4$. But it does not, $w_2(b_6) = 1/4$ whilst $w_1(b_6) = 3/16$, contradiction! ■

7.2 Generalizing $R_{\mathbb{A}}$

Recall that the idea behind the procedure leading to the minimisation of ambiguity was that if agents are asked to choose from two indistinguishable sets of options they should end up making similarly indistinguishable choices. In the characterisation given above in chapter 5, we defined indistinguishability by means of suitable permutations of A. However, we could push this notion further to capture the fact that the choice context, and specifically the choice problem at hand, does not change essentially if 0's and 1's are uniformly transposed.

The generalisation of the Minimum Ambiguity Reason that arises from this has an interesting application in the construction of a solution concept for pure coordination games and consequently a tight connection with radical interpretation problems. As to the former the idea is that of producing a Reason that facilitate the selection of so-called "focal points" among the strategies available in a coordination game. Allowing for a more general notion of indistinguishability here has the consequence of producing a more "context independent" notion of a focal point than it would be possible under the selection of strategies allowed by the canonical Minimum Ambiguity Reason described in chapter 5. As to the latter, this more general notion of indistinguishability permits us to capture a more basic notion of synonymity, a key concept intervening in the process of triangulation which underlie radical interpretation. These aspects will be fully developed in chapter 8 below, where the connections between Rationality-as-conformity, coordination problems and radical interpretation is discussed.

The key notion intervening in the generalization of $R_{\mathbb{A}}$ is that of a *transformation*.

Definition. *An injective function $j : K \to 2^A$ is 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 $j(K) = K$.*

Note that in what follows σ, σ' etc will always denote permutations of A and similarly for δ, δ' etc..

As for permutations above, transformations are closed under inverses and composition.

Lemma 7.3. *Let $j_1 : K \to 2^A$ be a transformation of K and $j_2 : j_1(K) \to 2^A$ a transformation of $j_1(K)$. Then $j_1^{-1} : j_1(K) \to K$ is a transformation of $j_1(K)$ and $j_2 j_1 : K \to j_2 j_1(K)$ is a transformation of K.*

Proof. Notice that if $j_1(f) = \delta_1 f \sigma_1$ and $j_2(h) = \delta_2 f \sigma_2$ for $f \in K$, $h \in j_1(K)$ then $j_1^{-1}(h) = \delta_1^{-1} h \sigma_1^{-1}$ for $h \in j_1(K)$ and $j_2 j_1(f) = \delta_2 \delta_1 f \sigma_1 \sigma_2$ for $f \in K$. ■

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 figure 3.1

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

Figure 7.1: 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 $K \in \mathbb{K}$, and j be a transformation of K. Then

$$j(R(K)) = R(j(K)).$$

Like Renaming, 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.

Clearly the Transformation Principle implies the Renaming Principle, just take δ to be the identity. Notice however that by the proof of Proposition 4.3 if we replace Renaming Principle by the Transformation Principle in Theorem 4.2 then the only Regulative Reason is the trivial one.

We can now define the procedure for the minimization of ambiguity in an entirely analogous way as before.

Definition. *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:*

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

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

This *generalized ambiguity construction* leads to a Reason which satisfies exactly the same properties which hold for the Minimum Ambiguity Reason discussed above in chapter 5. Indeed, the notion of transformation introduced here amounts to the automorphisms of a certain structure, exactly in the same as way the permutations before, except for the fact that in the present structure 0 and 1 are no longer distinguished elements. Hence, both constructions can be seen as special cases of a 'general ambiguity construction'. The fact that the Minimum Ambiguity construction (and the Smallest Uniquely Definable Reason) are so generalizable would appear to be one advantage that they have over the Regulative Reasons.

In particular we have the following generalization.

Theorem 7.4. *$R_{\mathbb{A}}$ satisfies Transformation.*

Proof. (In order to avoid an excess of parentheses we shall sometimes, as here, write jB etc. rather than $j(B)$ for $B \subseteq 2^A$.) As usual let j be a transformation of K. We

need to prove that

$$j(R_{\mathbb{A}}(K)) = R_{\mathbb{A}}(j(K)).$$

We first show by induction on m that for all $f \in K$, $j\mathbb{S}'_m(K, f) = \mathbb{S}'_m(j(K), j(f))$. So, for the base case, we want to show that j preserves the $\mathbb{S}'_0$-ambiguity classes, that is to say, for all $f \in K$,

$$j\mathbb{S}'_0(K, f) = \mathbb{S}'_0(j(K), j(f)).$$

Let

$$\mathbb{S}'_0(K, f) = \{g_1, \ldots, g_q\}.$$

Choose a transformation j' such that $j'(K) = K$ and $j'(f) = g_i$. Then $jj'j^{-1}$ is a transformation of $j(K)$ and $jj'j^{-1}(j(f)) = j(g_i)$. Hence, $j\mathbb{S}'_0(K, f) \subseteq \mathbb{S}'_0(j(K), j(f))$. Similarly, $j^{-1}\mathbb{S}'_0(j(K), j(f)) \subseteq \mathbb{S}'_0(K, f)$, so equality must hold here.

Assume now the result for the $\mathbb{S}'_m$-th ambiguity class, so we want to prove that

$$j\mathbb{S}'_{m+1}(K, f) = \mathbb{S}'_{m+1}(j(K), j(f)).$$

We distinguish between two cases, corresponding to the ones appearing in the construction of the ambiguity classes. Recall that $\mathbb{S}'_{m+1}(K, f) = \mathbb{S}'_m(K, f)$ if $m + 1 > |\mathbb{S}'_m(K, f)|$. So, in this case, the result follows immediately by the inductive hypothesis. Otherwise, since j is $1 - 1$, it is enough to see that

$$\begin{aligned} j\mathbb{S}'_{m+1}(K, f) &= j\{g \in K \mid |\mathbb{S}'_m(K, f)| = |\mathbb{S}'_m(K, g)|\} \\ &= \{j(g) \in j(K) \mid |\mathbb{S}'_m(j(K), j(f))| = |\mathbb{S}'_m(j(K), j(g))|\} \text{ (i.h.)} \\ &= \mathbb{S}'_{m+1}(j(K), j(f)). \end{aligned}$$

Since, by Proposition 5.3, $R_{\mathbb{A}}(K)$ is the smallest $\mathbb{S}(K, f)$, this concludes the proof of the Proposition. ■

Likewise we can prove a generalization of Theorem 5.5 above.

Theorem 7.5. *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 $R(K) = K'$.*

Proof. The direction from right to left follows immediately from the Transformation principle. For the other direction define, for $K_1 \subseteq 2^A, K_1 \neq \emptyset$,

$$R(K_1) = \begin{cases} j(K') & \text{if } K_1 = j(K) \text{ for some transformation } j \text{ of } K; \\ K_1 & \text{otherwise .} \end{cases} \tag{7.4}$$

Note that in the first case $R(K_1)$ is defined unambiguously, that is to say, whenever we have two transformations j_1, j_2 of K such that $K_1 = j_1(K) = j_2(K)$, then $j_1(K') = j_2(K')$. This follows since in this case, $j_1^{-1}j_2$ is a transformation of K and $j_1^{-1}j_2(K) = K$ so $j_1^{-1}j_2(K') = K'$, i.e. $j_1(K') = j_2(K')$.

We want now to show that if j is a transformation of K_1 to K_2, then $R(K_2) = jR(K_1)$. If K_1 is covered by the first case of (7.4), then so is K_2, for if j' is a transformation of K and $K_1 = j'(K)$, then $K_2 = j(j'(K))$ and jj' is a transformation of K by Lemma 5.1. In this case, $R(jK_1) = R(K_2) = jj'(K') = jR(K_1)$. If K_1 is covered by the second case of (7.4), so is K_2 since if $K_2 = j'(K)$ for some transformation j' of K, then $K_1 = j^{-1}j'(K)$ so $R(K_1)$ would be defined by the first case. It follows then that here we must have $R(j(K_1)) = R(K_2) = j(K_1) = jR(K_1)$ as required. ■

Note that we can generalise in exactly the same fashion the Smallest Uniquely Definable Reason. In this case we just have to delete 0 and 1 from the distinguished elements of the structure $\mathcal{M}$ thus obtaining the structure

$$\mathcal{M}' = \langle \{0,1\} \cup A \cup K, \{0,1\}, A, K, =, Comp \rangle.$$

It is natural to ask, at this point, what the Regulative Reasons would look like if we considered transformations in place of permutations there. However with this change the requirement of Renaming cannot be strengthened to what is expected here, i.e.

$$\delta R(K)\sigma = R(\delta K \sigma)$$

without reducing the possible Regulative Reasons to the trivial one alone – as can be seen by considering the initial step in the proof of Theorem 4.2.

Chapter 8

Focal points, triangulation and conformity

ABSTRACT: *We investigate two interconnected problems of "rational choice" - the selection of focal points in pure coordination games and the triangulation process in radical interpretation - in the light of the Rationality-as-conformity framework.*

Take again the Robotic Rendez-vous example introduced above in section 1.4.1 where two robotic agents I and II which have lost communication, need to meet at a certain location to restore it. In what is perhaps the simplest situation, any location is as good as any other, provided that I and II conform on it. How should the robots reason so as to achieve their goal? That is, how should they *choose* a location $\mathtt{l}$?

There is a close connection between *pure coordination problems* of this sort and situations of *radical interpretation*. After all, and very schematically, what I and II must do in order meet at an otherwise arbitrary location 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) choose accordingly. More specifically, once the possible locations say $\mathtt{l}_1, \ldots, \mathtt{l}_k$ are identified, agents should choose on the basis of some introspection in which they interpret each other by relating themselves to the "external world" – the representation of the choice problem. Since I and II

do not share a language, in fact they cannot communicate, the problem they end up facing is one of *radical* interpretation (recalled below).

At the same time this situation embeds the essential features of *strategic interaction*: what corresponds to the "rational" or "commonsensical" or even "logical" or 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 provide us with somehow precise and well-understood guidelines for the mathematical solution of our problem. In fact, as it will be shortly illustrated, the framework of non-cooperative games *does* provide us with a very clear and compact representation of the corresponding choice situation. Yet, as we noted above in section 2.3, as far as games of pure coordination are concerned, the classical solution concept based on Nash-equilibrium is of no use whatsoever, theoretically or practically.

The purpose of this chapter is to investigate this connection between pure coordination games and radical interpretation problems in the light of Rationality-as-conformity. In order to do this we shall idealise on the nature of the interpretation problem that we are to consider. At such a level of abstraction we shall be able to highlight how radical interpretation and pure coordination share a common structure. This latter is, in turn, extremely close to the Rationality-as-conformity main problem. Hence, the goal of our analysis will be to consider coordination games as well as radical interpretation problems in terms of Rationality-as-conformity. Our conclusion will be that the Minimum Ambiguity construction provides both a "natural" solution concept for pure coordination games based on the selection of "focal points" *and* the core of a procedure to initiate, from scratch, radical interpretation.

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), and Camerer points to "language" as "an obvious example" of coordination (Camerer, 2003, ch. 7), we have no knowledge of any attempt to relate mathematically the structure of *pure coordination* games with that of *radical*

interpretation.

Since coordination games do not correspond to the usual framework for discussing radical interpretation problems, we shall start by showing how, under suitable abstraction, the two situations reduce to a common structure.

8.1 A first example: radical translation

Some of the key aspects of radical interpretation received their first systematic discussion under the heading of radical *translation*. 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 prior knowledge. This complete lack of information, together with the fact that the two agents are assumed not to share a third language, makes the translation problem *radical*.

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

There are many subtleties connected with this example, none of which is of particular relevance for our present purposes. However the following issues involved in the radical translation problem are central to our discussion:

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

2. How far can she 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 the English language, I would utter RABBIT in those circumstances in which the native speaker uttered GAVAVAI". Conditionals of this form are 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 are inclined to adopt "similar" linguistic behaviours under "similar circumstances". Elevated to the status of a normative maxim, this is known as the *Principle of charity*.

The latter question relates to the fundamental *indeterminacy* of radical translation. Quine argues that there cannot be a *unique* translation manual which the linguist may be able to construct. Rather there must be a plurality of *equally acceptable* manuals, that is to say, equally supported by the available evidence. Yet the linguist can and should aim at reducing this indeterminacy by applying the Principle of charity throughout. In this way she would be lead to *discard* 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 unique translation manual may simply be underdetermined by the empirical evidence available to the translator.

We can note already at this stage how deep is the analogy between the guidance offered by the Principle of charity in the selection of a translation manual and the approach to rational choice based on Reasons (discussed above in section 3.1.2) which constitutes the backbone of the Rationality-as-conformity framework.

8.2 Triangulation in radical interpretation

The issue of radical translation and its relation with the Principle of charity are taken a step further by Davidson's investigations on *radical interpretation* (Davidson, 1984). Davidson assumes that the two agents involved in radical interpretation, despite being individually "rational" and willing to establish communication, do not happen to share any language. (Compare this with the discussion of the possible objections in the supermarket example of 1.1.) This can be the case of an adult and an infant who try to establish communication.

Note that the problem is more general than radical translation in two respects. Firstly, it applies to those situations, as the one just mentioned, in which an agent

(the infant) might potentially lack any language whatsoever. Secondly, besides being "foreign" as translation is, interpretation can as well be "domestic" (Davidson, 1984, p.125). Indeed one can press on (as Davidson does) and argue that given (an appropriate version of) the Inaccessibility assumption (cf. section 1.3) all interpretation is radical. For if nothing is known about another agent's "mental states", there can never be certainty about the fact that two speakers are actually using the *same* language.

The next important difference between Quine and Davidson's take on the problem is that in the context of radical interpretation, the Principle of charity is 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, and so on.

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 so-called "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 - as we have already pointed out - 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 portion of the external environment shared by the agents. It is the location of those causes that results in getting a first 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 generalization 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), p. 119)

Triangulation, hence, is form of conformity where the two agents aim at "converging" on the same interpretation of their linguistic behaviour. A fundamental aspect of the triangulation process consists in the recognition of the role played by constraints imposed by the "external world" on the interpretational choices. In particular, as a consequence of the Principle of charity, 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.

We now can appreciate how close analogues of the Fundamental assumption, as well as Likemindedness, Common knowledge and Saliency underlying the Rationality-as-conformity framework (see 1.3 above) play a fundamental role in the radical interpretation problem as well. This clearly encourages us to investigate radical interpretation from the Rationality-as-conformity point of view. In particular we wish to study the contribution of the latter towards providing a *procedure* to facilitate triangulation. In this attempt, however, we should be aware of the fact that if we were to consider the "full" case of linguistic interpretation, that is to say interpretation as performed by human beings in everyday situations (involving natural languages), we would run into enormous difficulties. Just to take an example, we should face the daunting task of providing a rigorous definition 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). Complications of this sort surely contribute towards the fact that our current understanding of radical interpretation doesn't seem include any clear-cut *procedure* by means of which agents can start triangulating.

In order to bypass those complications, we shall adopt here the 'mathematician's point of view' underlying the Rationality-as-conformity characterisations. Hence we shall abstract from the complications related to the use of natural language and the actual observation of non-verbal behaviour, which admittedly play an important role in the general account of radical interpretation among humans. As a consequence we shall consider radical interpretation in the context of a one-shot, pure coordination game. Within this framework we argue that the Minimum Ambiguity construction (especially in the generalised form of chapter 7), by contributing towards a general understanding of focal points in pure coordination games, allows us to formalise the key choice process intervening in triangulation. In fact, given its recursive nature, the Minimum Ambiguity Reason provides us with effective procedure to initiate triangulation.

It goes without saying that the structure within which this kind of solution is

provided 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.

8.3 The conformity game

Recall from section 3.1.1 above that *possible worlds* are all the maps from a finite set A to the binary set $2 = \{0, 1\}$. Nothing else is assumed about the structure of the set A. The *conformity game* is a two-person, non-cooperative, non-zero-sum game whose normal form goes like this: the domain of the game is $\wp^+(2^A)$, the set of non empty-subsets of possible worlds; players I and II are to choose one strategy from an element K of $\wp^+(2^A)$, identical for both agents up to permutations of A and 2 (more on this below). Hence each strategy available to the agents corresponds to one element of $K = \{f_1, \ldots, f_k\}$, say. Players get a positive payoff p if they play the same strategy and nothing otherwise, all this being common knowledge. (Figure 8.1 represents the conformity game for $k = 3$.)

		Player II		
		f_1	f_2	f_3
Player I	f_1	p, p	$0, 0$	$0, 0$
	f_2	$0, 0$	p, p	$0, 0$
	f_3	$0, 0$	$0, 0$	p, p

Figure 8.1: The conformity game.

Note that for present purposes we limit ourselves to case in which each identical pair of strategies yields a unique positive payoff p, so that any point in the diagonal is "as good as any other" as far as the agents are concerned: all that matters is that they conform on their choice. Hence, as usual for *pure coordination games*, the conformity game admits of multiple Nash-equilibria. And, as noted informally in section 2.3 above, in this case the traditional theory of non-cooperative games fails to be of

substantial help: if "rational choice" is based on the notion of a Nash-equilibrium, the players of a conformity game end up choosing randomly all the time.

Rationality-as-conformity, being process-based, helps us overcoming this difficulty. Recall that the key elements intervening in the representation of the conformity problem are 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. 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*. Taken from this angle, the problem of devising a solution concept for pure coordination games (and hence, for initiating triangulation) amounts to constructing a procedure to isolate, within a given set of strategies, the salient ones.

Consider again, 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 8.2.

	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 8.2: A representation on the strategy set K

It is immediate to see 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 to look for properties other than utility in order to characterise some of the options as those which are likely to be selected by another agent. At the same time, however, we want rule out the possibility that agents might 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 8.2) which agrees with 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. Notice that this puts us in the situation of the generalised Minimum Ambiguity construction of chapter 7. This greater generality, where 0 and 1 are not taken as distinguished elements of a certain structure, seems to fit better the intuition according to which focal points - both in the context of triangulation and in that of coordination games - must 'arise' from the structure of the choice problem while minimising the import of the assumptions on the other agents. Of course even greater generality could be introduced by dropping the assumption that agents face essentially similar (up to transformations) sets of strategies, namely by letting them "guess" which options their fellow players may be actually facing. This route, pursued for instance by Kraus et al. (2000), would however introduce a number of complications which, at least at this very first stage of formalisation, we prefer to avoid.

The conformity game meets the structure of radical interpretation to the extent that this latter is (i) abstracted to the case in which interpretation is defined over possible worlds, rather than the full natural language, and (ii) restricted to the process that enables triangulation. Hence, in particular, subsequent adjustments of triangulation that exploit the agents' capability of observing each others' non-linguistic behaviour (over time) fall beyond the scope of the conformity game. This captures the intuition that *radical* interpretation is somehow an intrinsically one-shot situation: once agents can rely on past experience or "confirmed hypotheses", the Inaccessibility assumption clearly ceases to hold.

8.4 From triangulation to focal points (and back)

The need for incorporating focal points in the game theoretic toolbox was firstly put forward by Schelling (1960). A key message of this monograph can be roughly summed up in the idea that whenever players *do not* have competitive interests, then the traditional, outcome-based solution concepts fall short of providing satisfactory accounts of rational choice. The specific focus of Schelling's investigation concerns the already recalled "tacit coordination" games with "common interest".

The fundamental feature of those games is their complete symmetry with respect to both players and strategies. If also equilibria are symmetric, coordination games are said to be *pure*. This makes utility-based solution concepts inapplicable and Schelling stressed this by referring to pure coordination games as "clueless" or "genius-proof". Rather, for players involved in such games

> [w]hat 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, p.54)

This passage makes the connections between pure coordination games, radical interpretation problems and Rationality-as-conformity extremely clear. On the one hand we can see that Schelling advocates for a triangulation-like solution for coordination games, where the "convergence" is about the mutual expectations – rather than the actual observation – of the other's behaviour. As discussed extensively above, this is what happens in the Rationality-as-conformity case. On the other hand, we can see that our Fundamental assumption underlies Schelling's characterisation of the problem, whilst in the context of radical interpretation this assumption is embedded in the Principle of charity.

The intuition underlying the solution based on focal points is that these correspond to strategies (courses of actions) which enjoy some degree of "saliency" or "conspicuousness", in Schelling's phraseology, which will lead agents to distinguish among options. A variety of perspectives on what saliency can be taken to be has been proposed in the literature (see, e.g. Sugden (1995); Janssen (1998); Kraus et al. (2000)). The Rationality-as-conformity framework suggests we consider saliency as arising from the 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. Mehta et al. (1994) refer to this latter as *Schelling's salience*.

On the basis of the empirical evidence obtained from controlled experiments, the authors argue that when faced with coordination problems akin to the conformity game, players behave as if they adopted the following two-steps process. Firstly, agents consider the rules (Reasons, in our terminology) that could be applied and then choose to adopt a rule which, if followed by their fellow players, would eventually facilitate conformity.

This explanation of the use of focal points in solving coordination problems is consistent with the approach to the selection of Reasons discussed above in section 5.4. There we stressed that the effectiveness (towards achieving conformity) of each of our Reasons depends essentially on the particular choice context, with the consequence that no Reason, justified as it may be, is likely to be optimal under any circumstances. Hence, we suggested, agents might consider several Reasons in turn, and choose to apply the one returning the subset of the initial set of possible worlds K with the smallest cardinality. In the context of Rationality-as-conformity this amount exactly to increasing one's chances to achieve conformity. Of course this solution is open to the obvious criticism that there might be situations in which distinct Reasons might yield subsets of K of the same cardinality.

The study of Schelling salience, hence, amounts to the study of focal points as identified through appropriate choice processes. The most distinctive constrains imposed on such a process turn out to be almost unanimously taken to be a combination

of *uniqueness* and *obviousness*. The idea being that uniqueness and obviousness would make a certain subset of possible worlds (strategies in the conformity game) *stand out* when considered in the context of the choice context faced the agents. In this sense the robotic rovers of our example would have good reasons to choose a location $\mathtt{l}_\mathtt{j}$ which stands out in the set $\{\mathtt{l}_1, \ldots, \mathtt{l}_k\}$. Naturally, if I can conclude that the location $\mathtt{l}_\mathtt{j}$ 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 $\mathtt{l}_\mathtt{j}$ 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 Kraus et al. (2000) suggested, this becomes a major concern once we take into account the limitations (i.e. bounded reasoning capabilities) of the agents. In other words, although a unique choice might be available to the agents, the computational expenses required to reach it would make following that path undesirable. Moreover, there could be circumstances in which appeal to uniqueness may lead to undesirable conclusions, as we had already occasion to remark when introducing the Minimum Ambiguity Reason. In particular uniqueness should never be pursued at the expense of running the risk of never agreeing on the final choice. In fact radical translation and interpretation warn us, as confirmed by the ubiquitous possibility sub-optimal Reasons, against aiming at uniqueness. Therefore it seems more appropriate, in general, to speak of aiming at facilitating coordination through focal points, in the same way we speak of facilitating conformity by adopting Reasons.

Given the inapplicability of the outcome-based solutions, in order to locate focal points and hence facilitate coordination (and hence triangulation) we need to introduce some *asymmetries* among the strategies available to the players of the conformity game. The structure of the problem makes the *Minimum Ambiguity Reason* introduced in above in chapter 5 a natural candidate for achieving this goal.

Our first argument is that - somehow by 'definition' - the Minimum Ambiguity

Reason aims at identifying those elements of a choice context K which are "naturally distinguished" within the structure of the choice problem. This meets the 'obviousness' requirement for focal points, namely the fact that must stand out. At the same time the Minimum Ambiguity construction aims at selecting the smallest subset of outstanding options of K, meeting the idea of 'uniqueness'. However in the presence of indistinguishable options, strict uniqueness should be abandoned as required by the Transformation principle. The upshot of this in the context of radical interpretation is of indubitable importance. In our abstract framework, in fact, the equivalence of possible worlds under transformations can be taken to capture the relation of *synonymity* among linguistic expression. Now if the interpretational choices of an agent were not closed under transformation, that is if any pair of indistinguishable possible worlds were not included among the set of best options from K, the resulting interpretation process would fail to reflect synonymity among expressions. But this would he highly undesirable as one of the ideal goals of translation as well as interpretation, in fact, consists in individuating systematically synonymy among linguistic expressions.

Surely the distinct levels of abstraction stand out in the comparison of the radical interpretation and the conformity game situations. While the radical interpretation problem is crucial in the attempt 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. For instance, an example of sure interest is the so-called *solution of anaphora* in linguistics which involves choosing the name to which a certain pronoun refers within a sentence like "the boy stared at the man. He beat him".

8.5 Concluding remarks

Summing up, radical interpretation (when restricted to the structure of the process of triangulation of mutual expectations) and pure coordination, can be seen as two faces of the same problem of "rational choice". Rationality-as-conformity, on the other hand, provides a unitary framework for investigating a solution concept for such a problem based on the minimisation of the ambiguity of possible worlds. As well as accounting for the use of focal points in pure coordination games the Minimum Ambiguity construction provides us with a procedure for initiating triangulation.

Note that in both cases the Minimum Ambiguity construction seems to be more adequate than the alternative Reasons provided by the Rationality-as-conformity framework. Interpretation involves choice among possible meanings, and this choice must be a disambiguating one. As to pure coordination games, the very notion of a focal point requires that these latter stand out within the context of a given choice problem, making the Minimum Ambiguity Reason a natural candidate for such choice processes.

As we have stressed, many of the investigations that followed Schelling's original intuitions can be seen as attempts at providing an explanation for the ability that human agents seem to have in exploiting focal points in order to achieve, or at least facilitate coordination. There has been a widespread scepticism, however, concerning the possibility of providing a mathematical solution to coordination games. Schelling himself, for instance, noted that

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

An entirely similar attitude is shared (four 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

> This *selection* problem is unsolved by analytical theory and will only be solved by observation. Camerer (2003)

Within the scope of its abstractions, Rationality-as-conformity counters this pessimistic view by pointing to a general mathematical solution to the problem of facilitating coordination where focal points are located through the application of Reasons. Among the Reasons investigated here, however, the Minimum Ambiguity construction seems to be able to place a serious bid for the most adequate choice process to select focal points, especially given its built-in bias towards favouring outstanding and "unique" choices.

That this very construction provides at the same time a procedure to initiate triangulation in (abstract and pre-linguistic) radical interpretation problems is, in our own view, a most intriguing connection.

8.6 Further remarks

It is interesting to compare, if briefly, the solution concept for the conformity game based on the Minimum Ambiguity construction with the solution concepts for coordination games arising from considerations of (approximate) common knowledge. An early, influential example on the role of 'structural' common knowledge was introduced with the *electronic mail game* (Rubinstein, 1989). The upshot of this investigation is that if it is assumed that players have common knowledge of the payoff structure, it turns out that they can coordinate efficiently (in terms of payoff) whereas if this assumption is weakened to *almost common knowledge* such an efficiency is lost. In other words, if common knowledge is replaced by a "high" but finite number of levels of knowledge in a way that at some depth $n+1$ one player will be uncertain about whether the other player has knowledge of depth n, the selection of the equilibrium for coordination can be forced to be inefficient. This situation is usually adduced as an example illustrating the qualitative difference between common knowledge and finite levels of knowledge.

An alternative take on this sort questions relating to "common knowledge", and specifically the role that higher order beliefs and mutual expectations have in the selection of multiple equilibria in coordination games, is given by the so-called *global*

games. The intuition here is that each player observes the payoff structure of the game "with a very small amount of noise" (Morris, 2002). This can be compared with the conformity game where each player is informed that she is receiving a strategy set with agrees only up to transformations with the one observed by the other player. And in fact, as we have seen, the intuition underlying construction on the Minimum Ambiguity Reason turns out to be closely related to "working out the higher-order beliefs that the information structure generates" (Morris, 2002).

A precise comparison between the conformity game on the one hand and global games and related situations on the other surely deserves deeper investigations.

Chapter 9

Summary and conclusions

We have introduced Rationality-as-conformity as a simple and abstract mathematical framework to investigate the notion of "rational choice". The central problem we have been focusing on is that of characterising the choice processes that two like-minded yet inaccessible agents might adopt in order to conform on the selection of a possible option. Three such choice processes have been discussed and their reciprocal connections investigated.

Rationality-as-conformity has been compared with some of the key mathematical accounts of "rational choice". First of all, we have pointed out how Rationality-as-conformity, in particular through the Regulative Reason, can be taken to provide a general justification for the Paris-Vencovská characterisation of probabilistic common sense. We have also shown, however, that a complete embedding of the latter in the Rationality-as-conformity framework seems not to be possible in the case in which possible worlds are interpreted probabilistically. A full explanation of such a failure surely deserve deeper investigations.

As to the more traditional accounts of rational choice, we have pointed out how many aspects of classical decision theory, game theory and social choice theory can be accommodated within our framework. Savage's characterisation is preserved in its spirit, though Rationality-as-conformity permits of a more general notion of distinguishability among possible options. In particular, the fact that possible worlds are not evaluated only in terms of their (expected) utility gives rise to a framework

for the study of rationality which allows us to overcome, without almost any effort, the difficulties presented by those game theoretic situations where strategies are not distinguishable on the grounds of their utilities. This process-based analysis of rational choice is very close to the spirit of the choice-functional approach to social choice theory.

The structure of Rationality-as-conformity resonates with important epistemological questions. We have illustrated this by recalling the problem of radical interpretation, which by virtue of its "primitive" character constitutes an analogue to our basic choice problem. In particular we have illustrated how a suitable abstraction of radical interpretation can be captured and given a (partial) algorithmic solution with the Minimum Ambiguity construction. Given that we have characterised rational choice in terms of the selection of the "obvious", "outstanding" or "logical" option, a related problem that appears to be amenable to investigation within the Rationality-as-conformity framework is the origin of linguistic convention.

9.1 Rationality-as-conformity as a logic

Although we have been using concepts and techniques from the mathematical logician tool-box, no consequence relation or proof system has been defined in the previous chapters. So, can we say that Rationality-as-conformity is a piece of logic?

Unsurprisingly this all depends on what a logic is taken to be. In the sense, which is perhaps the original one, of "what *the thinking* agent does" (Gabbay et al., 2002, p.12), Rationality-as-conformity is no doubt a logic. In particular it can be seen as a logic of practical reasoning.

> [...] a logic - a logic of practical reasoning, for example - gives a good account of itself to the extent to which it is able to define procedures for the competent production of practical reasoning. (Gabbay et al., 2002, p.8)

In fact we have noted on several occasions throughout the thesis, that Rationality-as-conformity accounts for what we often refer to in ordinary speech as the "logical thing to do".

It is an interesting exercise, at this point, to look at the main problem of Rationality-as-conformity from the other way round. When producing matrices (choice contexts) of the sort illustrated above, one often has the feeling that some choices would surely look more natural than others. Explaining this feeling requires a form of abduction which Rationality-as-conformity, in its three characterisations does provide. Future research along these lines may exploit this abductive framework to understand the origin of spontaneous conventions, that is, loosely, those things on which people find themselves to agree "for no particular reason".

Hence, we seem to be back to a point which we discussed at the end of chapter 4. There we pointed to a certain tension between thinking that the principles characterising the Regulative Reason can be adopted upon reflection, and the fact that when choosing according to one of the (non-trivial) Regulatives agents were not to be seen as consciously feeling an obligation to satisfy such principles. The abductive view of Rationality-as-conformity helps dissipating further this tension. In fact one could think of the act of choosing according, say to R_1 as a mainly unconscious business, whereas the formal explanation of such an act requires conscious evaluation of the principles involved. This opening to the sub-conscious elaboration of logical reasoning is explicitly taken into account in the discussion of *practical logic* (see, e.g. Gabbay et al., 2002; Gabbay and Woods, 2001), and surely resonates with the focal points approach to pure coordination problems discussed above.

9.2 Pluralism in Reasons

In proposing our solution for the main Rationality-as-conformity problem, we have in fact introduced what amount to four working Reasons, $R_0, R_1, R_{\mathbb{A}}, R_{\mathbb{U}}$. These arose through very different considerations. In the case of the Regulative Reasons through an adherence to rules, for $R_{\mathbb{A}}$ through an algorithm based on repeatedly trying to

fulfill two desiderata, and for $R_{\mathbb{U}}$ through picking the smallest uniquely definable set within the given structure of the problem. This plurality of approaches and answers raises a vexing question. How can we feel any confidence that there are not other approaches which will lead to entirely different answers?

As we have noted above, ideas and concepts from game theory would seem to have very definite application in generating Reasons. Furthermore similar hopes might be extended to other areas traditionally concerned with the formalisation of "rational choice", for example decision theory and social choice theory. Moreover other areas of mathematics might also lead to the production of Reasons.

The idea behind the construction of the Minimum Ambiguity Reason, for instance, is closely related to the study of certain permutational actions on symmetric groups, investigated at the borders of group theory and combinatorics. In particular the Minimum Ambiguity construction (especially in the generalised version) bears a close connection with the so-called Polyà theory of counting, where the problem investigated consists in computing the "essentially distinct" orbits induced by a permutational action (like, e.g. a transformation j) on a given symmetric group. Given the similarity of the problems addressed, there might well be an interesting and "new" way of generating Reasons within this area.

Analogous points could be made about model theory, with its interests in definable subsets, and Kolmogorov complexity, with its emphasis on minimum description length. In short, the answer to the vexing question is that we can have little such confidence beyond the modicum which comes from having failed to find any ourselves.

In fact, even with the candidates we already do have, we have seen that both the Regulative and Minimum Ambiguity Reasons appear capable, on their day, of monopolizing what could be perceived as the right, 'logical' answer. Hence one way of explaining this situation is that even in this very simple context (let alone in the real world) we should be ready to admit a plurality of good reasons giving rise to a plurality of (distinct) 'rational' arguments, rather than looking restlessly for a unique, universal one.

There seems to be a lesson for the supermarket shelving business too here.

Bibliography

J. Aczel. *Lectures on functional equations and their applications.* Academic Press, Mathematics in Science and Engineering, 1966.

M. Aizerman and A. Malishevski. General theory of best variants choice: Some aspects. *IEEE Transactions on Automatic Control*, 26:1030–1040, 1981.

R.C. Arkin. *Behavior-based robotics.* MIT Press, 1998.

C. Camerer. *Behavioral Game Theory: Experiments on Strategic Interaction.* Princeton, 2003.

F. Chu and J. Halpern. Great expectations. Part I. On the customizability of generalized expected utility. In *International Joint Conference on Artificial Intelligence*, 2003.

G. Coletti and R. Scozzafava. *Probabilistic Logic in a Coherent Setting.* Kluwer, Dordrecth, 2002.

R. Cox. Probability frequency and reasonable expectation. *American Journal of Physics*, 42:1–13, 1946.

D. Davidson. Radical Interpretation. In *Inquiries into Truth and Interpretation*, pages 125–140. Oxford University Press, 1984.

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

B. de Finetti. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17:289–329, 1931.

B. de Finetti. *Probability, Induction and Statistics*. Wiley, New York, 1972.

B. de Finetti. *Theory of Probability*. Wiley, New York, 1974.

B. de Finetti. *Filosofia della probabilità*. Il Saggiatore, Milano, 1995. English translation forthcoming.

D.M. Gabbay, R.H. Johnson, H.J. Ohlbach, and J. Woods. *Handbook of the Logic of Argument and Inference*. North-Holland, 2002.

D.M. Gabbay and J. Woods. The new logic. *Logic Journal of the IGPL*, 9(1):157–190, 2001.

H. Glock. *Quine and Davidson on language, thought and reality*. Cambridge University Press, 2003.

J. Halpern. *Reasoning About Uncertainty*. MIT Press, 2003.

J. Halpern, R. Fagin, Y.Moses, and M.Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

P.J. Hammond. Rationality in economics. *Rivista internazionale di scienze sociali*, 105:247–288, 1997.

R. Hilpinen. Carnap's new system of inductive logic. *Synthese*, 25:307–333, 1973.

J. Hintikka. *Knowledge and Belief: An Introduction to the logic of the two notions*. Cornell University Press, Ithaca, 1962.

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

M. Jaeger. Measure selection: Notions of rationality and representation independence. In G.F. Cooper and S. Moral, editors, *Proceedings of the 14th conference on Uncertainty in Artificial Intelligence, Madison, Wisconsin*, pages 274–281, 1998.

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

E.T. Jaynes. Where do we stand on Maximum Entropy. In D. Levin and M. Tribus, editors, *The Maximum Entropy Formalism*. Cambridge University Press, 1979.

G. Kalai, A. Rubinstein, and R. Spiegler. Rationalizing choice functions by multiple rationales. *Econometrica*, 70(6):2481–2488, 2002.

J.G. Kemeny. Fair bets and inductive probabilities. *Journal of Symbolic Logic*, 20 (3):1–28, 1955.

J.M. Keynes. *The General Theory of Employment Interest and Money*. McMillan, London, (1936), 1951.

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

T.A.F. Kuipers. Confirmation theory. In E. Craig, editor, *Routledge Encyclopedia of Philosophy*. Routledge, 1998.

D. Lehmann. Nonmonotonic logic and semantic. *Journal of Logic and Computation*, 11(2):229–256, 2001.

D. Lewis. *Convention: A philosophical study*. Harvard University Press, 1969.

R.D. Luce and H. Raiffa. *Games and Decisions*. Wiley, NY, 1957.

D. Marker. *Model theory: An Introduction*. Graduate Texts in Mathematics 217. Springer, 2002.

K. May. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, 20(4):680–684, 1952.

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

S. Morris. Coordination, communication, and common knowledge: a retrospective on the electronic-mail game. *Oxford Review of Economic Policy*, 18(4):433–445, 2002.

J.F. Nash. The bargaining problem. *Econometrica*, (28):155–162, 1950.

J.F. Nash. Non-cooperative games. *Annals of mathematics*, (54):286–295, 1951.

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

M.J. Osborne and A. Rubinstein. *A course in Game Theory*. MIT Press, Cambridge Massachussets, 1994.

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

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

J.B Paris. Common sense and maximum entropy. *Synthese*, 117(1):73–93, 1999.

J.B. Paris. A note on the dutch book method. In *Proceedings of the Second International Symposium on Imprecise Probabilities and Their applications*, Ithaca, NY, US, 2001.

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

J.B. Paris and A. Vencovská. In defence of the maximum entropy inference process. *International Journal of Approximate Reasoning*, 17:77–103, 1997.

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

C.R. Plott. Path independence, rationality and social choice. *Econometrica*, 41(6): 1075–1091, 1973.

W.V. Quine. *Word and Object*. MIT Press, Cambridge, Massachusetts, 1960.

A. Ram, R.C. Arkin, K. Moorman, and R.J Clark. Case-based reactive navigation: a method for on-line selection and adaptation of reactive robotic control parameters. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 27(3):376–394, 1997.

F.P. Ramsey. Truth and probability (1931). In Jr. H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*, pages 61–92. Wiley, New York, 1964.

H. Rott. *Change, Choice and Inference : A Study of Belief Revision and Nonmonotonic Reasoning*. Oxford University Press, 2001.

A. Rubinstein. The electronic mail game: strategic behavior under almost common knowledge. *American Economic Review*, 79:59–89, 1989.

A. Rubinstein. Comments on the interpretation of Game Theory. *Econometrica*, 59 (4):909–924, 1991.

L. Savage. *The Foundations of statistics*. Wiley, New York, 1954.

T. Schelling. *The strategy of conflict*. Harvard University Press, 1960.

A. Sen. Social choice theory. In K. Arrow and M. Intriligator, editors, *Handbook of Mathematical Economics*, volume 3, pages 1073–1181. Elsevier, 1986.

A. Sen. Maximization and the act of choice. *Econometrica*, 65(4):745–779, 1997.

A. Shimony. Coherence and the axioms of confirmation. *Journal of Symbolic Logic*, 20(3):263–273, 1955.

H.A. Simon. Rationality in psychology and economics. *Journal of Business*, 57(2): 209–224, 1986.

R. Sugden. Rational choice: A survey of contributions from economics and philosophy. *The Economic Journal*, 101(407):751–785, 1991.

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

P. Suppes. A Bayesian approach to the paradoxes of confirmation. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 198–207. Amsterdam: North-Holland, 1966.

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

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, New Jersey, 1944.