The Veil of Public Ignorance[1]

Klaus Nehring
Department of Economics
University of California, Davis
e-mail: kdnehring@ucdavis.edu

February 27, 2003

[1]An earlier version of this paper was circulated under the title "Utilitarian Cooperation under Incomplete Information". I thank participants at the conference on "50 years after Arrow's and Harsanyi's Theorems" at the Cowles foundation, seminar participants at the universities of Bielefeld, Bonn, Penn, Princeton, Rochester, and the Université de Quebec à Montréal, and Hans Gersbach, Friedrich Habermeier and Clemens Puppe for helpful comments.

Abstract

A theory of cooperative choice under incomplete information is developed in which agents possess private information at the time of contracting. It is assumed that the group of cooperating agents has agreed on a utilitarian "standard of evaluation" (group preference ordering) governing choices under complete information. The task is to extend this standard to choices whose consequences depend on the agents' private information. It is accomplished by formulating appropriate axioms of Bayesian coherence at the group level. Assuming the existence of a common prior, the first main result generalizes Harsanyi's (1955) classical characterization of utilitarian preference aggregation to incomplete information. The second main result shows that Bayesian coherence of group preferences is compatible with Interim Pareto Dominance only if a common prior exists. This generalizes and corrects the classical literature on consistent Bayesian preference aggregation under complete information: allowing for incompleteness of information, consistent Bayesian aggregation turns out to be possible even if agents' beliefs differ, as long as differences in beliefs can be attributed to differences in information.

1. INTRODUCTION

Traditionally, most theories of normative collective action are situated in contexts of certainty or assume at least "complete information", in the sense that agents know each others' beliefs. Yet it is by now a truism that asymmetries of information are pervasive, and that allocation and collective choice problems are often shaped decisively by them. In many cases, informational asymmetries exist already at the time when basic allocation decisions are to be made; such situations of "incomplete information" will occupy center stage of the current paper.[1] Examples abound: voters, in deciding on a voting rule (constitution), typically possess private information about their preferences already. So do couples, when compromising under conflicts of interest. Also, when a regulatory process is to be designed, the stakeholders (firms, customers, workers, recipients of externalities) possess private information about its costs and benefits, of the availability of alternative technologies and outside options, etc.. Whenever an incomplete contract needs to be renegotiated, incompleteness of information is likely to be an issue.

In this paper, we shall conceptualize collective choice in terms of a group of agents who jointly choose an allocation on the basis of some agreed-upon "criterion of group optimality" or "standard of evaluation"[2]. Incompleteness of information complicates collective choice in two basic ways. It implies first that a group of agents cannot choose directly a state-contingent allocation, but only an allocation mechanism that induces a state-contingent allocation via the equilibrium play of the agents. This leads to the familiar incentive compatibility constraints on the set of feasible allocations. Yet incompleteness of information entails an additional, much less studied difficulty: by its very nature, the *collective* choice of the allocation mechanism must be based on publicly available (commonly known) information only. Hence any underlying evaluation criterion must respect the group's ignorance of agents' private information; the group choice must be made, as it were, behind a *veil of public ignorance*. In particular, following Holmstrom-Myerson (1983) and others[3], the fundamental notion of Pareto efficiency needs to be formulated ex interim, at the stage when agents know their own information but not yet that of others; a state-contingent allocation f interim Pareto

[1] By contrast, when asymmetries of information arise only after the fundamental allocation decsions are taken, one speaks of "imperfect information".

[2] In the following, we shall use terms such as "group decision critierion", "standard of evaluation", "social ordering", "group preference" interchangeably.

[3] Independent originators of the notion of interim Pareto efficiency include Harris-Townsend (1981), Wilson (1978), and indeed, though lacking any discussion, Harsanyi-Selten (1972).

dominates another allocation g if it is *commonly known*[4] that every agent strictly prefers f to g.

In the following, we shall assume that the group has already accepted a social standard as the basis for collective choice under complete information, that is with commonly known probabilities; examples of such standards are the utilitarian and the Rawlsian (leximin) orderings, as well as the Nash bargaining solution. We shall ask how these standards are to be extended to situations of incomplete information. In the present paper, we shall concentrate on the case of a "utilitarian" criterion in the sense of Harsanyi's (1955) classic contribution. While controversial, especially in view of Diamond's (1967) celebrated criticism, Harsanyi-style utilitarianism is the central, most frequently used and best-behaved criterion of social optimality under uncertainty. It is also very much alive at a foundational level, as demonstrated, for example, by the recent contributions by Dhillon-Mertens (1999) and Segal (2000)[5]. The present focus on utilitarian standards allows one to zero in on the fundamental epistemic issues that arise form the nature of collective choice under incomplete information per se, for any standard. In future work, we plan to extend the analysis to non-utilitarian criteria, a task which raises a number of additional, potentially controversial normative issues.

To get a better grasp of the issues involved, consider a situation a la Myerson-Satterthwaite (1983) in which two risk-neutral agents have the opportunity to trade an indivisible good. Assume that the agents already know their own reservation price for the good before selecting a trading mechanism,[6] and that they agree on the utilitarian criterion under complete information, evaluating outcomes in terms of the sum of utilities in the money metric. If, counterfactually, the two agents were able to choose the mechanism before knowing their reservation price, ranking mechanisms would simply be a matter of computing the expected total gains from trade ex ante. By contrast, under incomplete information, when the agents know their reservation price *before* choosing the mechanism, it is no longer self-evident how to rank mechanisms. Specifically, it is not clear on what basis a social expectation can be taken, and, indeed, whether a social expectation can be taken at all, since ex interim both agents have different information and thus different beliefs; moreover, these beliefs are not commonly known, and thus simply not available as the basis for cooperative, hence public,

[4]An event E is commonly known if it is the case that E, that everyone knows that E,that everyone knows that everyone knows that E, etc. .

[5]The former can be viewed as yielding a particular normalization of individuals' utility functions summed by the utilitarian criterion. This is also holds for the special case of the latter contribution in which domains consist of a unique resource vector.

[6]This assumption makes sense of Myerson-Satterthwaite's crucial assumption of *ex-interim* participation constraints; I thank Urs Schweizer for pointing this out.

decision making. Heuristically, what is needed is a "group belief" that is derived from the agents' individual beliefs but relies on commonly known information only. Can such a "group belief" be meaningfully determined?

Constructing the Common Prior as the Group Belief

Whether and how a "group belief" can be determined depends in part on the structure of the agents' beliefs about each other. For beliefs that are consistent with the existence of a common prior, the first main result of the paper, Theorem 1, the "Aggregation Theorem", identifies the common prior as the searched-for group belief. This result effectively generalizes Harsanyi's (1955) classical result to incomplete information. Specifically, it yields social preferences of the form

$$E_\mu \sum_{i \in I} U_i^f,$$

where the random variable U_i^f is agent i's utility derived from "act" (mapping from states to consequences) f, and E_μ denotes the expectation with respect to the common prior μ. This representation is based on three underlying axioms: Interim Pareto Dominance, Utilitarianism under complete information, and State Independence; the latter in effect allows one to apply the utilitarian criterion state-by-state. Neither independence nor completeness is assumed for group preferences over general acts.

In the context of contract theory, the Aggregation Theorem provides a justification for applying Myerson-Satterthwaite's (1983) computation of the optimal trading mechanism ex interim to situations in which the selection of the trading mechanism itself takes place. Moreover, in risk-neutral settings without participation constraints a la D'Aspremont-Gerard-Varet (1979) and Arrow (1979) in which there exist mechanisms that guarantee ex-post efficient allocations, the Aggregation Theorem justifies choosing such mechanisms over the others as utilitarian optimal ex interim.[7]

[7]Note that there typically are many mechanism that fail to guarantee ex-post efficiency but are not dominated ex interim by some mechanism that does; the selection of ex-post efficient mechanisms thus cannot be justified on grounds of interim efficiency alone.

When is "Bayesian Aggregation" Possible under Incomplete Information ?

The Aggregation Theorem assumes that agents' beliefs admit a common prior. This is necessary: the second main result of the paper, Theorem 2, the "Possibility Theorem", shows that it is possible for social orderings to jointly satisfy State Independence and Interim Pareto Dominance *only* if agents' beliefs admit a common prior. This result generalizes and corrects the existing (im)possibility results on Bayesian aggregation in the literature, all of which have been formulated in the context of complete information (see especially Hylland-Zeckhauser 1979, Hammond 1981, Broome 1990, Seidenfeld-Kadane-Shervish 1989 and Mongin 1995); that literature concluded that group preferences can satisfy Bayesian coherence together with ordinary Pareto Dominance only if agents' beliefs coincide. By contrast, according to the Possibility Theorem, Bayesian coherence (State Independence) at the group level is compatible with respect for consensual preference as long as differences in beliefs among agents can be fully attributed to differences in information.[8] While empirically this assumption is frequently violated as people often "agree to disagree" (Aumann 1976), it is possible to argue that such disagreement cannot happen among intersubjectively rational agents. If this *normative* interpretation of the CPA is accepted, the Possibility Theorem is especially satisfying since then the possibility of obtaining a well-defined group belief through Paretian aggregation is predicated upon the (intersubjective) rationality of individual, much as expected utility maximization at the social level presupposes the same at the level of individuals.

The Possibility Theorem implies that in the absence of a common prior, either Bayesian coherence or the Interim Pareto condition needs to be given up. In work in progress (Nehring 2003b), we shall propose an evaluation criterion that maintains the latter while weakening Bayesian coherence appropriately.

Related Literature

While the literature on non-cooperative decision making under incomplete information is vast, cooperative decision making has received far less attention in this setting. The attention that it has received is almost exclusively in the positive, strategic vein; following the seminal paper by

[8] In the sense that the different beliefs can be viewed as updates on a common prior. This interpretation of the common prior assumption is often referred to as the "Harsanyi doctrine". Its meaningfulness in a static setting is controversial. While Dekel-Gul (1997) have raised doubts, Bonanno-Nehring (1999) and Nehring (2001) endowed it with content even in a static setting in which there is no dynamic notion of receiving information.

Wilson (1978), this literature focuses mainly on incomplete information versions of the core and its properties. By contrast, we are aware of only two prior contributions to the theory of *"norm-based"* cooperative decision making under incomplete information with a motivation broadly comparable to ours, namely the early papers by Harsanyi-Selten (1972) and Myerson (1984) on the two-person Nash-bargaining solution under incomplete information. Since both of these propose extensions of a different, non-utilitarian optimality criterion, we shall defer a more detailed discussion to future work[9], besides noting that while both assume interim Pareto efficiency, neither foregrounds the central epistemic issue of "group belief".[10]

Plan of the Paper

In section 2, we illustrate the general idea of "group choice behind the veil of public ignorance" through a simple voting example. Section 3 introduces type spaces first in a fairly standard way in terms of probabilistic beliefs, and then in a richer decision-theoretic version in terms of agents' (interim) preferences. The main results of the paper, the Aggregation and Possibility Theorems, are derived and discussed in sections 4 and 5, respectively. Section 6 concludes. All proofs can be found in the appendix.

2. A VOTING EXAMPLE

To illustrate some of the central issues and concepts, we shall briefly study a highly stylized model of single-issue voting. Voters $i \in I$ have quasi-linear preferences over the adoption of a "project" $y \in Y = \{0, 1\}$ and net transfers t_i $u_i(y, t_i) = y\theta_i + t_i$, where θ_i denotes i's benefit of the project. For maximal simplicity and not infrequent realism, we shall assume that transfers are de facto infeasible; their role is merely to measure voters' preference intensity. For the sake of comparison, we shall first describe the cooperative choice of a voting mechanism *ex ante,* i.e. under the assumption that voters have not yet received private information about their preferences. Thereafter, we shall look at the same cooperative choice problem *ex interim*, if the voting mechanism is to be chosen when voters already know their own preferences but not others.

[9]The working paper version (Nehring 2002) compares the role of state-independence assumptions and of the common prior in these two contributions to that in the present one.

[10]The lack of an emphasis on epistemic issues probably reflects the unavailability of crucial concepts at the time; indeed, Harsanyi-Selten (1972) predates even the fundamental concept of common knowledge (formalized first by Aumann 1976).

A state $\omega \in \Omega$ is a profile of net benefits $\theta = (\theta_i)_{i \in I}$. Ex ante, the agents have a common prior μ over $\Omega = \mathbf{R}^I$; they anticipate that ex interim, they will be informed of their own net benefit θ_i, and of nothing else. An agent's "type" can thus be identified with his net benefit θ_i. For simplicity, assume that types are independent, i.e. that $\mu = \prod_{i \in I} \mu_i$. Voters must choose an incentive-compatible (direct) mechanism $f \in \mathcal{F}^{ic}$, i.e. an "act" $f : \Omega \to \{0, 1\}$ that depends only the sign of voters' net benefit: $f(\theta) = \widehat{f}(\sigma(\theta))$ with $\sigma(\theta) = (sgn\ \theta_i)_{i \in I}$, and such that $\widehat{f}$ is non-decreasing in $\sigma(\theta)$; note that incentive-compatibility is equivalent here to strategy-proofness.

We need the following notation maintained throughout the paper. A random variable Z is a real-valued function on Ω; constant random-variables are denoted in bold-face; $E_\mu Z$ denotes the expectation of Z with respect to the common prior μ, $E_\mu Z = \sum_{\omega \in \Omega} \mu(\omega) Z(\omega)$. Agent i's expectation of the random variable Z, when viewed as a function of the state, is again a random variable $E_i Z$ given by $E_i^\alpha Z = \sum_{\omega \in \Omega} p_i^\alpha(\omega) Z(\omega)$.

Mechanism Selection Ex Ante

In deciding on an optimal voting mechanism, voters have agreed upon a utilitarian standard given by the summation of utilities in the money metric. Ex ante, its meaning is conceptually unproblematic: it amounts to choice of a mechanism $f \in \mathcal{F}^{ic}$ that maximizes $E_\mu \sum_{i \in I} U_i^f$, where $U_i^f(\theta) := u_i(f(\theta), 0)$.

Observation 1 *An optimal mechanism is given by*

$$f^*(\theta) := \begin{cases} 1 & \text{if } \sum_{i \in I} E_{\mu_i}[\theta_i \mid sgn\ \theta_i] > 0 \\ 0 & \text{otherwise} \end{cases}.$$

The optimal mechanism f^* is simply determined by maximizing $\sum_{i \in I} u_i(y)$ given knowledge of $\sigma(\theta)$. It can be interpreted as a voting rule in which a vote by voter i *for* the project has weight $E_{\mu_i}[\theta_i \mid sgn\ \theta_i > 0]$, while a votes *against* the project has weight $E_{\mu_i}[-\theta_i \mid sgn\ \theta_i < 0]$. Thus, if the expected intensity of a positive preference exceeds that of a negative preference, a voters' "pro" vote will weigh more heavily than his "con" vote. Moreover, if voters' expected intensities in both directions are equal, different voters' weights in the optimal mechanism will reflect their ex-ante expected strength of preference $E_{\mu_i}[|\theta_i| \mid sgn\ \theta_i]$.

To simplify even further, assume that the common prior is symmetric in voters, i.e. that $\mu_i = \mu_*$ for all $i \in I$. Then the optimal mechanism must be anonymous, and thus amount to a "voting by quota" or "supermajority" rule. The voting-by-quota mechanism f_q (with quota q) is defined by

setting

$$f_q(\theta) := \begin{cases} 1 & \text{if } \frac{\#\{i|\theta_i>0\}}{\#\{i|\theta_i\neq 0\}} > q, \\ 0 & \text{otherwise} \end{cases}.$$

Observation 2 *With a symmetric common prior, voting by quota is optimal.*

The optimal quota q^* is equal to $\frac{E_{\mu_*}[|\theta_i||\theta_i<0]}{E_{\mu_*}[|\theta_i||\theta_i<0]+E_{\mu_*}[|\theta_i||\theta_i>0]}$.

Typically, the optimal quota will differ from 0.5; for example, if pro voters care more than con voters in expectation, it will take less than 50 % of pro-votes for the project to be accepted.[11]

With a symmetric common prior, it is also possible to justify the optimal mechanism by an efficiency argument. Indeed, ex ante, every agent prefers the quota q^* to any other, since for all $i \in I$ and $q \in [0,1] : E_\mu U_i^{f_{q^*}} \geq E_\mu U_i^{f_q}$. This leads immediately to the following observation.

Observation 3 *f_{q^*} is the unique anonymous mechanism that is ex-ante efficient among all incentive-compatible mechanisms $f \in \mathcal{F}^{ic}$.*

Mechanism Selection Ex Interim

Assume now that voters already know their own type (θ_i^τ) when selecting the voting mechanism (quota), maintaining the assumption of a symmetric common prior. Is the mechanism f_{q^*} still uniquely optimal ex-interim? And if so, in what sense? Does the argument from Pareto efficiency survive?

To address these questions, one needs to recognize first that under incomplete information, the relevant criterion of Pareto efficiency is that of interim Pareto efficiency due to Holmstrom-Myerson (1983): a mechanism f is *interim Pareto efficient* if there exists no other feasible mechanism g that is *commonly known* to be preferred by every agent, where each agent evaluates mechanisms based on his current (interim) beliefs. Formally, f **interim Pareto dominates** g if it is common knowledge that $E_i^\alpha U_i^f > E_i^\alpha U_i^g$ for all $i \in I$, where $E_i^\alpha U_i^f$ is i's interim expected utility under f. In our example, *any* quota is interim efficient.

Observation 4 *For any $q \in [0,1]$, f_q is interim Pareto efficient.*

To see this, consider a type with a positive net benefit $\theta_i > 0$. Such a type prefers mechanisms with a higher (interim) probability that the project will be accepted; this probability is given as

[11] A number of recent papers propose ex-ante optimizing accounts of supermajority rules, including Aghion et al. (2002), Gerardi and Yariv (2002), and Persico (2002).

$\prod_{j\neq i}\mu_j(\{\theta_{-i} \mid f(\theta_i,\theta_{-i})=1\})$. Evidently, this probability will be the larger, the smaller the quota q is; in particular, the quota $q=0$ is his most preferred, as it guarantees acceptance of the project. Conversely, any type with $\theta_i<0$ is interested in minimizing the chance that the project is accepted, with $q=1$ as the best. This shows that the mechanisms based on extreme quotas f_0 and f_1 are interim Pareto efficient; a slightly more involved argument establishes this also for intermediate quotas. The example illustrates a general phenomenon: typically, there are far more interim efficient that ex-ante efficient allocations.

The argument from Pareto efficiency thus ceases to work ex interim. On the other hand, the basic utilitarian intuition that the optimal mechanism should reflect the strength of preference indicated by a vote continues to apply. How can this intuition be captured? A straightforward but naive approach would be to maximize $\sum_{i\in I} E_i^\tau U_i^f$, where E_i^τ is agent i's expectation operator based on his true type θ_i^τ. The problem with this criterion is simply its inapplicability: no one knows its value, since every voter knows only the value of his own term in the sum. Intuitively, a viable criterion must in some manner "abstract" from agents' *private* expectations and extract a "publicly accessible version" of them to arrive at a decision criterion that can serve as the basis for collective decision making "*behind the veil of public ignorance*".

With independent types, the following heuristic consideration shows how this can be achieved. The key is the observation that while ex-interim the value of $E_iU_i^f$ is known only to voter i, all *other* voters have the same interim belief about i's type (given by the commonly known probability measure μ). Thus others' expectations $E_jE_iU_i^f$ of $E_iU_i^f$ are commonly known; being independent of j, they can be written as $E_{\neq i}E_iU_i^f$. It follows that a viable interim utilitarian criterion is given by

$$\sum_{i\in I} E_{\neq i}E_iU_i^f.$$

This criterion evaluates mechanisms on the basis of the sum of agents' interim expected utilities, *as estimated by the others.* It is easily seen that in fact

$$\sum_{i\in I} E_{\neq i}E_iU_i^f = \sum_{i\in I} E_\mu E_iU_i^f = E_\mu\left(\sum_{i\in I} U_i^f\right). \tag{1}$$

Thus, in the special case of independent types, we have obtained a heuristic rationale for using the common prior as the basis for an interim group expectation. In particular, we have provided an intuition why the common prior may be relevant and usable even though it does not coincide with anyone's belief ex interim. As shown at the end of section 4, this heuristic interpretation can be

extended to the non-independent case. Note that under the "interim utilitarian" criterion (1) yields the same optimal voting rule in every state, since what is commonly known does not depend here on the state. All information aggregation happens *within* the optimal mechanism, rather than in the choice of the optimal mechanism itself.

We have not yet argued why the common prior is *the* right group probability to use, nor have we justified the expectational functional form of the social ordering. The missing argument will take the form of a decision-theoretic axiomatization. To prepare the ground, we need to introduce a decision-theoretic formulation of type spaces. To ease the reader into it, this will be preceded by a review/reformulation of standard Bayesian type spaces.

3. FORMAL FRAMEWORK

3.1. Representing Incomplete Information by Type Spaces

Agents' mutual uncertainty will be modelled in terms in terms of type spaces. For reasons of both technical convenience and conceptual transparency, we define type spaces slightly differently from the usual by obtaining types as a derived construct. Specifically, we shall define a type space simply as a state space in which at any state α the agents' beliefs at that state p_i^α are specified. In addition, since we want to model an "interim" perspective on which agents already know their own private beliefs, a particular state is formally singled out to describe agents' actual beliefs.

Definition 1 *A* **rooted type space** *is a tuple* $\langle I, \Omega, \{p_i\}_{i \in I}, \tau \rangle$*, where*

- I *is a finite set of* agents.
- Ω *is a finite set of* states*; the subsets of* Ω *are called* events.
- *for every agent* $i \in I$, p_i *is a function that specifies, for each state* $\alpha \in \Omega$, *his probabilistic beliefs* $p_i^\alpha : 2^\Omega \to \mathbf{R}$ *at* α.
- $\tau \in \Omega$ *is the true state.*

Note that states occur twice in the representation of agents' belief, with $p_i^\alpha(\{\omega\})$ denoting agent i's probability in state α that state ω occurs. Since the states in a type space describe all agent's beliefs at that state, an agents' belief describes not only his beliefs about facts of nature, but also his

beliefs about other agents' (first-order) beliefs about states of nature. For example, the expression $p_i^\alpha(\{\omega|p_j^\omega(rain) \geq 0.7\})$ denotes agent $i's$ probability at state α that agent j believes that it will rain with at least 70% probability. This can be iterated; hence an agent's beliefs at a state specify his beliefs about agents' higher-order beliefs about states of nature, thus in effect: an entire belief hierarchy. Indeed, a state in a type space can simply be thought of as a notational device for describing the belief hierarchies of each agent.[12] Fixing a particular state τ as the "root" fixes a particular profile of belief hierarchies.

We will maintain the following two assumptions.

Assumption 1 *(Introspection)* For all $\alpha \in \Omega$ and all $i \in I$: $p_i^\alpha(\{\omega \in \Omega \mid p_i^\omega = p_i^\alpha\}) = 1$.

Assumption 2 *(Truth)* For all $\alpha \in \Omega$ and all $i \in I$: $p_i^\alpha(\{\alpha\}) > 0$.

Introspection says that agents are always (at any state α) certain of their own belief p_i^α. The Truth assumption states that, at any state that might occur, agents put positive probability on that state; agents therefore can never be never wrong in their probability-one beliefs. While standard, this assumption is not unrestrictive.[13]

For any $\alpha \in \Omega$, i's type at state α is defined formally as the set of states the agent thinks possible, $T_i(\alpha) := \{\omega \in \Omega \mid p_i^\alpha(\omega) > 0\}$. By Introspection and Truth, the family $T_i := \{T_i(\omega) \mid \omega \in \Omega\}$ is a partition of Ω, i's *type partition*. Note that defined types defined in this manner can be understood in the usual way: agents' beliefs are determined by their type, and agents always know their own type.

The conventional interpretation of a type space is dynamic, describing a point in time (the "interim stage") at which agents have updated their prior beliefs upon receiving some private information signal, and where agents' prior beliefs had been commonly known at an "ex-ante" stage. This *dynamic* interpretation is appealed to in the standard narrative accompanying asymmetric information models, and the reader may assume it for the sake of familiarity. In formal terms, a dynamic interpretation assumes as primitives agents' priors q_i (with support Ω) and information partitions T_i; an agents' type corresponds to the *signal received* $T_i(\alpha)$, and the interim beliefs are derived as *conditional* probabilities, $p_i^\alpha = q_i(./T_i(\alpha))$.

[12] By results due to Armbruster-Boege (1979) and Mertens-Zamir (1985), any profile of probabilistic belief hierarchies has a type-space representation; the assumption that the state space Ω is finite is restrictive but entirely standard. Infinite state-spaces are considered in Halpern (1998) and Feinberg (2000).

[13] See Bonanno-Nehring (1999) for a detailed study of its relaxation in the context of rooted type spaces.

On closer reflection, assuming the existence of a prior stage at which beliefs were commonly known seems highly restrictive and implausible: in many situations in which agents' have private information about their own preferences or abilities at a given point in time, they *always* knew more about their own preferences or abilities than others, hence the posited prior stage never existed.[14] In the absence of such a prior stage, interim beliefs are *unconditional* probabilities describing agents' mutual uncertainty about each other. This *static* interpretation will be the "official" one adopted throughout the paper; nonetheless, we will typically use the dynamic ex-interim/ex-ante terminology due to its suggestiveness and entrenchment.

An agent "knows" an event E at α if he is certain of it, i.e. if $E \supseteq T_i(\alpha)$. Let T_I denote the finest common coarsening of the partitions $\{T_i\}_{i \in I}$, with $T_I(\alpha)$ denoting the cell of T_I containing state α. E is *common knowledge* if everybody knows that E, and if everybody knows that everybody knows that E, and so forth. Formally, E is "common knowledge" at α if $E \supseteq T_I(\alpha)$. Since type spaces serve as a notational vehicle to represent hierarchies of beliefs, we will assume throughout that the state space includes only states that are relevant to their description, i.e. that $T_I = \{\Omega\}$. It is then unambiguous to speak of "common knowledge of an event", without reference to the state.

A probability measure $\mu : 2^\Omega \to \mathbf{R}$ is a **common prior** if, for all $i \in I$, $\omega \in \Omega$ and $A \subseteq \Omega$, $p_i^\omega(A) = \mu^\alpha(A/T_i(\omega))$ whenever $\mu(\omega) > 0$. In view of the partitional structure of the T_i and the assumption that Ω is the only common knowledge event, it is easily verified that if a common prior exists, it is unique, hence commonly known, and assigns positive probability to every state.

This completes the epistemic part of the model.

3.2. Utilitarian Type Spaces

We will now enrich this set-up by describing agents in terms of their (mutually uncertain) preferences over state-contingent outcomes ("acts"); from these, agents' beliefs and utilities can be derived.

Let X be a set of deterministic social alternatives, with typical element x. Let $\mathcal{L}$ denote the set of probability distributions on X, with typical element ℓ; to avoid technicalities, we shall confine attention to the set of "simple lotteries", that is: probability distributions with a finite number of outcomes with positive probability. In the manner of Anscombe-Aumann (1963), an *act* f maps states to probability distributions of social alternatives, $f : \Omega \longrightarrow \mathcal{L}$. In a multi-agent version

[14] Some authors have recently gone further and argued that postulating a prior complete information stage may not be meaningful even as a counterfactual; see in particular Dekel-Gul (1997) and Gul (1998).

of Anscombe-Aumann's "horse race" interpretation, a state $\omega \in \Omega$ describes the outcome of the horse race together with a profile of agents' belief hierarchies over that outcome. A social act is given by conducting a state-contingent "roulette lottery" which selects the social alternative x with conditional probability $f_x(\omega)$. It is understood that this conditional probability distribution is shared among all agents. Let $\mathcal{F}$ denote the set of all such acts, with $\mathcal{F}^{const}$ as the subset of constant acts; these correspond to the playing of the same lottery in every state, and will thus typically be referred to by the name of that lottery $\ell \in \mathcal{L}$.

Since individuals are mutually uncertain about each others' beliefs, they must be mutually uncertain about each others' preferences over acts, too; formally, these are random variables $\alpha \longmapsto \succeq_i^\alpha$, where $\succeq_i^\alpha$ is a preference relation on $\mathcal{F}$. Agents are commonly known to be expected utility maximizers with subjective probability measure p_i^α at state α and von Neumann-Morgenstern utility function $u_i : X \to \mathbf{R}$. To streamline notation, we shall write $u_i(\ell) = \sum_{x \in X} \ell_x u_i(x)$, and define, for any act f and agent i, a random variable U_i^f by setting $U_i^f(\omega) = u_i(f(\omega))$; U_i^f describes agent i's expected utility under f conditional on state ω.

Since the social ordering is derived from individual preferences, in principle it too is a random variable $\alpha \longmapsto \succeq_I^\alpha$. However, to serve as a basis for collective, hence public action, the social ordering itself needs to be public, that is: commonly known among agents; since we have assumed Ω to be the unique commonly known event, the social ordering can be treated as a constant $\succeq_I$. Here, we shall assume that the group already has accepted a "utilitarian" decision criterion $\succeq_{I|\mathcal{L}}$ for situations of complete information with agreed-upon probabilities. That is, the group ranks lotteries according to $\sum_{i \in I} u_i(\ell)$.

All of this is summarized in the following formal definition of the model of the paper.

Definition 2 *A* **utilitarian type space** *is a tuple* $\langle I, \Omega, \{\succeq_i\}_{i \in I}, \succeq_{I|\mathcal{L}}, \tau \rangle$*, where*

1. *I is a finite set of agents;*
2. *Ω is a finite set of states, with τ denoting the actual state;*
3. *For every agent $i \in I$, $\succeq_i$ is a mapping that specifies, for each state $\alpha \in \Omega$, a preference relation $\succeq_i^\alpha$ over acts, and $\succeq_{I|\mathcal{L}}$ is a social ordering over lotteries such that there exist mappings $\{p_i\}_{i \in I}$ and utility functions $\{u_i\}_{i \in I}$ such that*
   (a) *the $\{p_i\}_{i \in I}$ satisfy Introspection and Truth;*

(b) for all $f, g \in \mathcal{F}$, all $i \in I$ and all $\alpha \in \Omega$:

$$f \succeq_i^\alpha g \text{ if and only if } E_i^\alpha U_i^f \geq E_i^\alpha U_i^g;$$

(c) for all tuples $(\overline{u}_i)_{i \in I}$, there exists a lottery $\ell \in \mathcal{L}$ such that $u_i(\ell) = \overline{u}_i$ for all $i \in I$;

(d) for all $\ell, \ell' \in \mathcal{L}$:

$$\ell \succeq_{I|\mathcal{L}} \ell' \text{ if and only if } \sum_{i \in I} u_i(\ell) \geq \sum_{i \in I} u_i(\ell').$$

Utilitarian type spaces can be defined directly in terms of conditions on the primitives exploiting the representation theorems of Anscombe-Aumann (1963) and Harsanyi (1955) in a straightforward manner; see the working paper version Nehring (2002) for details.[15] Note that due to the representation requirements b) and d), utilities are unique up to addition of constants, agent by agent, and multiplication by a *common*, strictly positive factor. By a), the tuple $\langle I, \Omega, \{p_i\}_{i \in I}, \tau \rangle$ is a rooted type space as defined above. Assuming the u_i to be independent of the state amounts to assuming agents preferences over lotteries $\succeq_{i|\mathcal{L}}^\alpha$ to be commonly known. This assumption is made for simplicity and can be substantially weakened. The domain richness assumption c) is also substantially stronger than necessary; however, if it is given up, then the rationality assumptions on the social ordering $\succeq_I$ need to be strengthened. For a detailed discussion of both of these points, see Nehring (2002) again.

4. UTILITARIANISM EX INTERIM

The goal is to try to use the publicly available information about agents' beliefs (implicit in their preferences) to derive from the given standard $\succeq_{I|\mathcal{L}}$ comparisons of general "intersubjectively uncertain" acts f. Technically, we are looking for an **extension** $\succeq_I$ of $\succeq_{I|\mathcal{L}}$, that is, for a transitive and continuous[16] super-relation of $\succeq_{I|\mathcal{L}}$.

[15] Here, Harsanyi's (1955) Theorem has the role of helping to *characterize,* that is: to define, utilitarian type spaces axiomatically. It need not be assigned the role of *grounding* them normatively. The substantive justification of a particular utilitarian standard in a particular situation might be derived from other considerations such as those of Harsanyi (1953), Dhillon-Mertens (1999) or Segal (2000). A particular standard might also be derived from a group "decision" to compare utility differences across individuals in a particular way, without underlying philosophical foundation. For example, and very roughly speaking, in a private-goods economy society may decide *by fiat* to count everyone's utility gain from moving from the poverty line to becoming a milloniare equally.

[16] Continuity is understood here to mean that the sets $\{g | g \succeq_I f\}$ and $\{g | g \preceq_I f\}$ are closed for all f in the topology of pointwise convergence.

Almost by definition, consensual group choice will respect agents' unanimous preference; under incomplete information, such unanimity must be public. This is leads to the axiom of "Interim Pareto Dominance."

Axiom 1 (Interim Pareto Dominance) *$f \succeq_I g$ (resp. $f \succ_I g$) whenever it is commonly known that $f \succeq_i^\alpha g$ (resp. $f \succ_i^\alpha g$) for all $i \in I$.*

We will also assume that the social ordering satisfies a minimum of "Bayesian rationality". Specifically, we will adapt a version of the standard State Independence axiom which entails that the utilitarian ranking $\succeq_{I|\mathcal{L}}$ can be employed state by state.

Axiom 2 (State Independence) *$f \succeq_I g$ whenever it is commonly known that $f(\omega) \succeq_{I|\mathcal{L}} g(\omega)$.*

If agents' beliefs admit a common prior, there exists a unique extension of the utilitarian standard $\succeq_{I|\mathcal{L}}$; this extension maximizes the expected sum of agents' utilities based on the common prior.

Theorem 1 (Aggregation Theorem) *Let $\langle I, \Omega, \{\succeq_i\}_{i\in I}, \succeq_{I|\mathcal{L}}, \tau\rangle$ be a utilitarian type space with a common prior μ. Then there exists a unique extension $\succeq_I$ of $\succeq_{I|\mathcal{L}}$ satisfying Interim Pareto Dominance and State Independence; for this extension, $f \succeq_I g$ if and only if*

$$E_\mu\left(\sum_{i\in I} U_i^f\right) \geq E_\mu\left(\sum_{i\in I} U_i^g\right). \tag{2}$$

Note how this result embodies the idea of group choice "behind the veil of public ignorance": both Interim Pareto Dominance and State Independence exploit commonly known information only; these allow to derive a uniquely determined, complete ranking of general acts that is commonly known itself by construction. Note also that neither the Independence axiom nor completeness of the social ordering is assumed. To understand the logic of the result, consider the following example.

Example 1. Assume that social alternatives are allocations of "money" to the agents ($X = \mathbf{R}^I$), that agents are risk-neutral in money with $u_i(x) = x_i$ for all $i \in I$, and that a utilitarian standard is given by $\sum_i u_i$. Let $I = \{1, 2\}$, $\Omega = \{\tau, \beta, \gamma, \delta\}$, with τ as the true state. Agents' beliefs are given by $T_1 = \{\{\tau, \beta\}, \{\gamma, \delta\}\}$, $T_2 = \{\{\tau, \gamma\}, \{\beta, \delta\}\}$, and $\mu = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ as the common prior.

In natural notation, let f denote the act $(-2, -2; -2, -2; -2, -2; 10, 10)$, g the act $\mathbf{0}$, and h the act $(-2, -2; 4, -8; -8, 4; 10, 10)$. Since $E_\mu\left(\sum_{i\in I} U_i^f\right) = 2 > E_\mu\left(\sum_{i\in I} U_i^g\right)$, by Theorem 1, it must be the case that $f \succ_I g$. This can indeed be derived from commonly known information using the auxiliary act h as follows. Note first that $E_1 h = E_2 h = (1, 1, 1, 1)$. Hence it is commonly known

that both agents prefer act h to act g; by Interim Pareto Dominance, h is thus socially preferred to g, $h \succ_I g$. On the other hand, the acts f and h have the same total income, hence the same sum of utilities, in each state. It is thus commonly known that $h(\omega)$ is socially as good as $f(\omega)$, whence by State Independence, $h \sim_I f$. By transitivity, this implies that $f \succ_I g$.

We now comment on further aspects of this result, focussing on the nature of information aggregation in social optimization, and on the relation between individual belief and the common prior as the group's as-if belief.

The Veil of Public Ignorance and the Role of Communication.—

Example 1 illustrates how agents can deduce the interim utilitarian ranking (2) from commonly known information. Combining this ranking with information about feasibility constraints, any agent can determine (by himself) the socially optimal act(s), in the voting example of section 2, for instance, the optimal quota rule. This act is then "selected" publicly, e.g. the optimal quota is made law. Note that *no information has been exchanged yet*; the collective choice of the mechanism has been made without peering through the veil of public information. Now the selected mechanism is run, agents vote, say, and a social alternative is chosen on the basis of the revealed private information.[17]

While this scenario establishes the conceptual coherence of collective choice behind the veil of public ignorance as choice without communication, one may still wonder whether the restriction to using commonly known information only is artificially limiting, and whether the group may do better by trying to use some of the agents private information *in the choice of the mechanism*. However, in view of the revelation principle, such doubts are misplaced. In particular, one must bear in mind that any such communication must be credible, that is: respect incentive-compatibility constraints. Moreover, any incentive-compatible and socially desirable communication will already be built into

[17]This procedure of determining the collective choice from common knowledge information can be viewed as part and parcel of the very notion of consensual group choice under incomplete information. By contrast, the relevant feasibility constraints on acts will depend on the particular situation. One scenario (indeed the standard scenario of mechanism design theory) involves an impartial mediator who a) has complete control over all communication between agents, and b) who knows all that is common knowledge among the group. By (a), the mediator can select *any* incentive-compatible, individually-rational mechanism; by (b), the mediator can deduce the optimal mechanism among them from his information and run it. In other scenarios, a wide variety of further feasibility constraints may be relevant; these may result, for example, from incomplete control over agents' communication, the possibility of renegotiation, group participation constraints in the manner of the core, considerations of contract simplicity etc..

the optimal mechanism. [18]

These points about the role of communication can be illuminated further by taking another look at Example 1. In this example, it may seem paradoxical that f is ranked above g at the true state τ, even though at τ both agents know that they are made worse off by f than by g. However, this fact is only known but not *common knowledge* among the two agents, and thus not available to the construction of the interim criterion $\succeq_I$.

Of course, this fact could be made available easily by communication among the two agents. To fix ideas, agents 1 and 2 might decide whether f or g is chosen by a simultaneous voting procedure in which "f" is be selected if and only if both agents vote "Yes". The only reasonable behavior in this game is for types $T_1(\tau)$ and $T_2(\tau)$ to vote "No", while the types $T_1(\delta)$ and $T_2(\delta)$ vote "Yes".[19] This selects f in state δ and g otherwise, thereby in effect implementing the act $e = (0, 0; 0, 0; 0, 0; 10, 10)$. Since e is ranked strictly above both f and g by the interim criterion $\succeq_I$, e is recommended by $\succeq_I$ ahead of f, resolving the apparent paradox, and illustrating the general point that any socially desirable and feasible communication will be realized by the optimal act.

The Interpretation of Common Priors.—

It is also worth commenting on the status of the common prior. Evidently, Theorem 1 nowhere assumes that individual agents "forget" information, or catapult themselves back to some actual or fictitious ex-ante stage, so as to make the common prior their own belief. This is a fundamental difference between the "veil of public ignorance" as understood here, and the traditional "veil of ignorance".[20] Here, the common prior is meaningful first of all at the group level as part of the representation of the interim utilitarian criterion (2).

Nonetheless, it is of natural interest to express this criterion in terms of the interim beliefs of individual agents. Based on a characterization of the common prior due to Samet (1998), this can be done using agents' higher-order expectations of the sum of utilities as follows: $f \succ_I g$ if and only

[18] It may, of course, simply happen sometimes that some agents try to "leak" some of their private information to others in the hope of influencing the group decision in a favorable manner; if they are successful, this just means that the informational basis for collective choice has changed. Whether or not successful "cheap talk" among agents is indeed possible in equilibrium depends on the situation and deserves further study; Holmstrom and Myerson (1983) discuss a related issue under the heading of "durability".

[19] The described behavior is the only one that survives two rounds of eliminating weakly dominated strategies.

[20] As suggested briefly is section 3 and argued more extensively by Gul (1998) and Dekel and Gul (1997), in the absence of an actual ex-ante stage, such thought-experiments may not be meaningful in principle.

if, for some finite sequence $\{i_1, ..., i_k\}$, it is common knowledge that

$$E_{i_k}...E_{i_1}\left(\sum_{i\in I} U_i^f\right) > E_{i_k}...E_{i_1}\left(\sum_{i\in I} U_i^g\right).$$

As illustrated by the voting example of section 2, in the special case of independent types, already the value of the second-order expectation E_jE_kZ is commonly known for any $j \neq k$ and Z, and equal to $E_\mu Z$. Hence in this case, the interim utilitarian standard can be defined in terms of second-order expectations, i.e. $f \succeq_I g$ if and only if

$$E_jE_k\left(\sum_{i\in I} U_i^f\right) \geq E_jE_k\left(\sum_{i\in I} U_i^g\right) \text{ , for any } j \neq k.$$

Group Choice versus Welfare Interpretation.—

In order to clarify the informational assumptions behind the preference and belief aggregation described by the Aggregation Theorem, we have interpreted the social ranking $\succeq_I$ as a decision criterion for group choice under incomplete information. Alternatively, the ordering $\succeq_I$ can also be viewed as a "social welfare ranking"; specifically, the Aggregation Theorem captures the spirit of the "*ex-ante* school" adapted to situations of incomplete information; the hallmark of the ex-ante school is the unrestricted acceptance of the Pareto principle applied to agents' preferences at the time of the group decision, and hence incorporating their current (here: their interim) beliefs.[21] By contrast, the "ex-post school" subscribes to the Pareto principle only state by state; it typically interprets the social ordering as that of an "impartial observer" or "benevolent dictator" with beliefs of his own. For the ex-post school, incompleteness of information among agents presents no new normative issues of interest, since, in any case, it is the observer's beliefs that count.

[21] These two interpretations are linked in that Paretian interim welfare comparisons are naturally interpreted in terms of *hypothetical* group choices, i.e. as determining what the agents *should* choose *if* the state-contingent allocation was a matter of collective choice (rather than, say, non-cooperative interaction). Indeed, under incomplete information, an interpretation of interim welfare judgements in terms of *some* well-defined (hypothetical) choice situation under uncertainty seems necessary to pin down their meaning unambiguously by establishing a well-defined informational basis; otherwise, it would simply not be clear from what beliefs these welfare judgments could be derived. On the group choice interpretation, this informational basis is the sum total of everything that is commonly known by the group.

5. WHEN IS BAYESIAN COHERENCE COMPATIBLE WITH THE PARETO PRINCIPLE ?

The Aggregation Theorem construes the extended social standard as subjective expected utility maximization at the social level; the common prior as the group's subjective probability measure can be viewed as an aggregating individual agents' belief from behind the "veil of public ignorance". If agents' beliefs fail to be consistent with a common prior, such aggregation is no longer possible. Not only does the common prior become unavailable as the canonical "consensus group belief", social expected utility maximization becomes incompatible with the Pareto principle. The source of the incompatibility is especially transparent in the risk-neutral case. Here, if agents' beliefs are inconsistent with a common prior, there exists a mutually profitable bet by the no-betting characterization of common priors due to Morris (1994) (cf. Theorem 3i) of the Appendix). Interim Pareto Dominance requires a group preference for this bet over not betting; by contrast, State Independence entails social indifference between the two, since the bet involves a mere reshuffling of money=utility among agents in each state, a matter of indifference under a utilitarian standard.

Theorem 2 (Possibility Theorem) *The utilitarian type space* $\langle I, \Omega, \{\succeq_i\}_{i\in I}, \succeq_{I|\mathcal{L}}, \tau\rangle$ *admits an extension* $\succeq_I$ *that satisfies Interim Pareto Dominance and State-Independence if and only if agents' beliefs are consistent with a common prior.*

Example 2. Assume agents to be risk-neutral etc. as in Example 1, and $I = \{1, 2\}$ and $\Omega = \{\tau, \beta, \gamma, \delta\}$. Let beliefs be given as follows: $p_1^\tau = p_1^\beta = (\frac{2}{3}, \frac{1}{3}, 0, 0)$, $p_1^\gamma = p_1^\delta = (0, 0, \frac{1}{3}, \frac{2}{3})$, $p_2^\tau = p_2^\gamma = (\frac{1}{3}, 0, \frac{2}{3}, 0)$, $p_1^\beta = p_1^\delta = (0, \frac{2}{3}, 0, \frac{1}{3})$. Since these beliefs do not admit a common prior, there exists a mutually advantageous bet. One such bet is $f' = (1, -1; -1, 1; -1, 1; 1, -1)$; indeed, it is common knowledge that $E_1 f' = E_2 f' = \frac{1}{3}$. Thus, by Interim Pareto Dominance, $f' \succ_I g = \mathbf{0}$. On the other hand, by State Independence, $f' \sim_I g$, the contradiction asserted by the Theorem.

Theorem 2 generalizes the existing (im)possibility results on Bayesian aggregation in the literature in the context of complete information; see Hylland-Zeckhauser (1979) and others quoted above. These concluded that Bayesian coherence is compatible with ordinary Pareto Dominance only if agents' beliefs are identical. By contrast, according to the Possibility Theorem , differences in beliefs among agents do not force a choice between Bayesian coherence and respect for consensual preference per se; the conflict arises only when these differences cannot be fully attributed to differences in information in the sense that the different beliefs can be viewed as updates on a common prior. The

Possibility Theorem thus removes the flair of "paradox" of the extant results; indeed, on a Harsanyian point of view on which agents' belief *should* admit a common prior as a matter of (intersubjective) rationality, it shows that Bayesian and Paretian aggregation is possible *whenever* agents are fully rational in this sense. [22] Conversely, it is not surprising that irrationalities at the individual level lead to a conflict between (Bayesian) rationality at the social level and the Pareto Principle, since the latter ties social preference closely to individual preference. In this regard, the role of the common prior assumption can be viewed as analogous to that of individual expected-utility maximization.

In the case of inconsistent beliefs, Theorem 2 forces a choice between Bayesian coherence or respect for consensual preference. The considerations relevant to making this choice are largely similar to those arising in the special case of complete information, where it has been extensively if inconclusively discussed in the literature. Both alternatives have found their advocates: Raiffa (1968, p. 233-237) argues for maintaining the Pareto principle in the context of group choice, even with heterogeneous beliefs, while others including Broome (1991), Mongin (1995,1998) and Gilboa-Samet-Schmeidler (2001) have argued in favor of Bayesian coherence. The appropriate choice clearly depends critically on the intended interpretation of the social aggregation. On the one hand, if the group itself is viewed as the seat of preference (as assumed in this paper), then it is hard to see what more compelling ground for group preference there could be than unanimity; on the other hand, if the social ordering $\succeq_I$ is that of a benevolent "outside observer" (philosopher king), then it is natural to strive for one unified socially relevant belief (the observer's), and unanimity among the members of the group themselves may not be viewed as a decisive ground for preference by the outsider. Indeed, the critics of the Pareto principle tend to assume the "outside observer" interpretation, most explicitly Mongin (op. cit.).[23]

On the Pareto Principle in the Presence of Inconsistent Beliefs.—

There is no need to discuss in detail the merits of the relevant arguments, with the potential exception of a recent argument by Gilboa-Samet-Schmeidler (2001) that has precursors in Levi (1981)

[22] In the present framework, the known impossibility results reappear if ordinary Pareto dominance is formulated as "joint improvement" at the true state as follows:

$$f \succeq_I^\tau g \text{ (resp. } f \succ_I^\tau g) \text{ whenever } f \succeq_i^\tau g \text{ (resp. } f \succ_i^\tau g) \text{ for all } i \in I.$$

[23] A particular variant of the outside observer interpretation is that of an "expert panel", in which the agents are "experts" making recommendations for an organization, whose judgments are to be aggregated by an "executive" on behalf of the organization; in this context, Raiffa (1968), too, favors Bayesian coherence over the Pareto principle.

and Mongin (1995). According to this argument, the unanimous preference of, for example, a bet that merely redistributes total utility such as the preference of f' over g in Example 2 is by itself not a compelling *ground* for a corresponding social preference of f' over g, since agents unanimous preference is based on conflicting reasons for such preference (different utility and probability assessments). If this argument is found compelling, one may conclude that the Possibility Theorem is conceptually uninteresting, on the grounds that one of its key premises is not well-founded. However, we shall argue that such a conclusion would be premature, since Interim Pareto Dominance can be derived from premises not subject to this criticism.

Specifically, consider the following weakening of Interim Pareto Dominance that we shall refer to as "Non-Paternalism". Non-Paternalism says that if all but one agent do not care which one of two acts is chosen, in that they are indifferent between the outcomes *in every state*, than group preference follows the preference of the agent to whom the comparison matters.

Axiom 3 (Non-Paternalism) *If, for some $i \in I$, it is common knowledge that $f \succeq_i^{\alpha} g$ (respectively $f \succ_i^{\alpha} g$), and that, for all $j \in I \backslash \{i\}$, $f(\alpha) \sim_j g(\alpha)$, then $f \succeq_I g$ respectively $f \succ_I g$.*

Note that if interpreted in terms of underlying beliefs as grounds for the social preference, any particular instance of Non-Paternalism appeals only to the beliefs of the one agent who cares, and is thus not subject to the Gilboa et al. critique. Nonetheless, using a transitivity argument, one can derive the general Pareto criterion from Non-Paternalism, as stated formally in the following observation.

Observation 5 *Non-Paternalism and Transitivity imply Interim Pareto Dominance.*

We illustrate the logic of the Observation in the context Example 2; its generalization is straightforward. To obtain the desired instance of Interim Pareto Dominance from Non-Paternalism, consider the act $h' = (1, 0; -1, 0; -1, 0; 1, 0)$. By Non-Paternalism, clearly $f' \succ_I h'$, since the choice between f' and h' matters only to the second agent. Likewise, since the choice between g and h' matters only to the first agent, $h' \succ_I g$. By transitivity of social preference therefore $f' \succ_I g$.

Thanks to Observation 5, Theorem 2 survives as a live, provocative "impossibility theorem" when Interim Pareto Dominance is replaced by Non-Paternalism in its statement. By consequence, when agents' beliefs are inconsistent, Bayesian coherence implies that the preferences of at least (one type of) one agent must be disrespected, "paternalized". Such disrespect is justified if the "social chooser" can lay legitimate claim to superior judgement, but seems problematic otherwise. In particular,

even if one subscribes to the CPA normatively as a condition of intersubjective rationality, it seems doubtful that a violation of the CPA provides in itself sufficient grounds for such disrespect. For such violation implies at most that *some* type of *some* agent is "irrational",[24] but there is nothing in the inconsistent belief system that would allow one to localize the irrationality in a particular type of a particular agent, and that would thereby justify overriding the preferences of this particular type. Thus, in the absence of appropriate additional information that would pinpoint particular types as deficient in rationality, Non-Paternalism maintains a powerful appeal even when beliefs are inconsistent.

We have focused on the common prior case in this paper partly in order to stay clear from these fairly subtle and controversial normative issues, in order to focus squarely on the issues that arise from incomplete information per se. While the CPA is clearly restrictive from the point of view of "raw empiricism", it is extremely widely used in economic models, and can therefore hardly be an altogether unreasonable idealization of reality. Moreover, any formal normative argument presupposes "idealizing" assumptions, such as, for example, the assumption that agents maximize expected utility, or, as in the existing literature, that agents have complete information.

In any case, one should not conclude from Theorem 2 that the very notion of group choice from behind a veil of public ignorance becomes unworkable when beliefs do not satisfy the CPA. In work in progress (Nehring 2003b), we propose an interim utilitarian criterion for general beliefs that ranks acts according to functionals of the form

$$E_\eta \left(\sum_{i \in I} E_i U_i^f \right),$$

where η is an appropriate "compromise prior" that agrees with the common prior in the consistent case.

6. CONCLUSION

The goal of this paper has been to demonstrate the conceptual coherence of the notion of "group choice from behind a veil of public ignorance"; it has been realized in the Aggregation Theorem of section 4, under the assumptions of a utilitarian standard and of a common prior. The key difference between a "veil of *public* ignorance" and the traditional "veil of ignorance" is the ex-interim perspective of the former that relies exclusively on agents' actual beliefs, in contrast to the

[24]For an account of the CPA that allows one to make this inference rigorous, see Nehring (1998).

ex-ante perspective of the latter that relies on a counterfactual thought-experiment. Under the assumptions of this paper, it turns out that group choice from behind a veil of public ignorance can be understood *as-if* from behind a classical veil of ignorance at some fictitious ex-ante stage in which all incompleteness of information has been removed. The Aggregation Theorem specifically delivers two things: first, the "group probability" governing group choice ex interim is the (commonly known) common prior. In addition, the group uses the same decision criterion ex-interim that it uses ex-ante, maximization of the common-prior expectation of the sum of agents' utilities. This ability to move to an ex-ante stage greatly simplifies the application of the interim utilitarian criterion in applications. Indeed, as we have shown, a number of existing results in the literature such as the optimization in Myerson-Satterthwaite (1983) can naturally be understood as applications of this criterion. An obvious challenge for future work is to explore the notion of group choice behind the veil of public ignorance under more general assumptions.

APPENDIX: PROOFS

The proofs of Theorems 1 and 2 rely on the following characterization of common priors and their existence that is naturally interpreted in behavioral, betting terms. The behavioral interpretation assumes the existence of a transferable currency with respect to which all agents are risk-neutral. The idea is to determine those random-variables (viewed as contingent payments to the group, "bets") which the group I would be willing to bet on collectively using an appropriate sharing arrangement; as part of the "rules of the game", it must be commonly known that the sharing arrangement is strictly acceptable to each agent. This is formalized in the following definition.

Definition 3 *f is* **acceptable for** *I if there exist $f_i : \Omega \to \mathbf{R}$ for $i \in I$ such that $f = \sum_{i \in I} f_i$ and such that it is common knowledge that $E_i^\alpha f_i > 0$ for all $i \in I$.*

Theorem 3 *i) A common prior exists if and only if $\mathbf{0}$ is not acceptable for I.*

ii) If a common prior μ exists, f is acceptable for I if and only if $E_\mu f > 0$.

While part i) is well-known (see Morris (1994)), part ii) is a novel result proved in a separate note, Nehring (2003a). The latter is used in the proof of the Aggregation Theorem, the former in that of the Possibility Theorem.

Proof of Theorem 1.

Necessity is straightforward.

To prove sufficiency, consider two acts $f, g \in \mathcal{F}$ such that $E_\mu \sum_i U_i^f > E_\mu \sum_i U_i^g$.

Let $Z = \sum_i U_i^f - \sum_i U_i^g$; by Theorem 3, ii) there exist $\{Z_i\}_{i \in I}$ such that $\sum_i Z_i = Z$ and such that it is common knowledge that $E_i Z_i > \mathbf{0}$. Hence by domain richness, there exists some act $h \in \mathcal{F}$ such that $U_i^h = U_i^g + Z_i$ for $i \in I$. By Interim Pareto Dominance, evidently

$$h \succ_I g.$$

On the other hand, since $\sum_i U_i^h = \sum_i U_i^g + \sum_i Z_i = \sum_i U_i^f$, $h(\omega) \sim_{I|\mathcal{L}} f(\omega)$ for all $\omega \in \Omega$, hence by State Independence

$$h \sim_I f.$$

By transitivity,

$$f \succ_I g.$$

By continuity, this also implies that $f \succeq_I g$ whenever $E_\mu \sum_i U_i^f \geq E_\mu \sum_i U_i^g$. $\square$

Proof of Theorem 2.

In view of Theorem 1, we only need to show necessity of the common prior.

Assume thus that agent's beliefs do not admit a common prior. By Theorem 3, i), there exists a vector $(Z_i)_{i\in I}$ such that a) $\sum_{i\in I} Z_i = 0$ and such that b) it is common knowledge that $E_i Z_i > 0$ for all $i \in I$. By domain Richness, there exist acts f and g in $\mathcal{F}$ such that $U_i^f = Z_i$ and $U_i^g = \mathbf{0}$ for all $i \in I$. The pair f and g establishes the desired contradiction in view of a) and b).

Indeed, by b), it is common knowledge that $f \succ_i^{\omega} g$. Hence by Interim Pareto Dominance

$$f \succ_I g.$$

On the other hand, by a) and the existence of a utilitarian standard represented by $\sum_i u_i$, $f(\omega) \sim_{I|\mathcal{L}} g(\omega)$. Hence by State Independence

$$f \sim_I g,$$

the desired contradiction. $\square$

REFERENCES

[1] Aghion, P., Alesina, A., and F. Trebbi (2002), Endogenous Political Institutions, mimeo, Harvard University.

[2] Anscombe, F. G. and R. Aumann (1963), A Definition of Subjective Probability, *Annals of Mathematical Statistics,* 34, 199-205.

[3] Armbruster, W. and W. Boege (1979), Bayesian Game Theory, in *Game Theory and Related Topics,* ed. by O. Moeschlin and D. Pallaschke, Amsterdam, North-Holland.

[4] Arrow, K. (1979), The Property Rights Doctrine and Demand Revelation under Incomplete Information, in M. Boskin (ed.), *Economics and Human Welfare,* Academic Press, New York.

[5] Aumann, R. (1976), Agreeing to Disagree, *Annals of Statistics,* 4, 1236-1239.

[6] Blackorby, C., Donaldson, D. and P. Mongin (2000), Social Aggregation Without the Expected Utility Hypothesis, mimeo.

[7] Blackorby, C., Donaldson, D. and J. Weymark (1999), Harsanyi's Social Aggregation Theorem for State-Contingent Alternatives, *Journal of Mathematical Economics,* 32, 365-387.

[8] Bonanno, G. and K. Nehring (1999), How to Make Sense of the Common Prior Assumption under Incomplete Information, *International Journal of Game Theory,* 28, 409-434.

[9] Broome, J. (1991), *Weighing Goods,* Basil Blackwell, Oxford.

[10] D'Aspremont, C. and L. A. Gerard-Varet (1979), Incentives and Incomplete Information, *Journal of Public Economics,* 11, 25-45.

[11] Dekel, E. and F. Gul (1997), Rationality and Knowledge in Game Theory, in: Kreps D. M. and K.F. Wallis (eds.), *Advances in Economics and Econometrics*, vol. 1, Cambridge, Cambridge UP, 87-172.

[12] Dhillon, A. and J.-F. Mertens (1999), Relative Utilitarianism, *Econometrica* 67, 471-498.

[13] Diamond, P. (1967), Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility: Comment", *Journal of Political Economy,* 75, 765-766.

[14] Gerardi, D. and L. Yariv (2002), Putting Your Ballot Where Your Mouth is, mimeo.

[15] Gilboa, I. , Samet, D. and D. Schmeidler (2001), Utilitarian Aggregation of Beliefs and Tastes, mimeo.

[16] Gul, F. (1998), A Comment on Aumann's View, *Econometrica,* 66, 923-927.

[17] Hammond, P. (1981), Ex Ante and Ex post Welfare Optimality under Uncertainty, *Economica* 48, 235-250.

[18] Harsanyi, J. (1953), Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking, *Journal of Political Economy,* 61, 434-435.

[19] Harsanyi, J. (1955), Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility, *Journal of Political Economy,* 63, 309-321.

[20] Harsanyi, J. (1967-68), Games with Incomplete Information Played by Bayesian Players, Parts I-III, *Management Science,* 8, 159-182, 320-334, 486-502.

[21] Harsanyi, J. and R. Selten (1972), A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information, *Management Science,* 18, P80-P106.

[22] Holmstrom, B. and R. Myerson (1983), Efficient and Durable Decision Rules with Incomplete Information, *Econometrica,* 51, 1799-1819.

[23] Hylland, A. and R. Zeckhauser (1979), The Impossibility of Bayesian Group Decision Making with Separate Aggregation of Beliefs and Values, *Econometrica,* 47, 1321-1336.

[24] Mertens, J.-F. and S. Zamir (1985), Formulation of Bayesian Analysis for Games with Incomplete Information, *International Journal of Game Theory,* 14, 1-29.

[25] Mongin, P. (1995), Consistent Bayesian Aggregation, *Journal of Economic Theory,* 66, 313-351.

[26] Mongin, P. (1998), The Paradox of the Bayesian Experts and State-Dependent Utility Theory, *Journal of Mathematical Economics,* 29, 331-361.

[27] Morris, S. (1994), Trade with Heterogeneous Prior Beliefs and Asymmetric Information, *Econometrica,* 62, 1327-1347.

[28] Myerson, R. (1984), Two-Person Bargaining Problems with Incomplete Information, *Econometrica*, 52, 461-487.

[29] Myerson, R. and M. Satterthwaite (1983), Efficient Mechanisms for Bilateral Trading, *Journal of Economic Theory,* 28, 265-281.

[30] Nehring, K. (1998), Common Priors for Likeminded Agents, mimeo UC Davis.*
(The *ed manuscripts are available at http://www.econ.ucdavis.edu/faculty/nehring/)

[31] Nehring, K. (2001), Common Priors under Incomplete Information: A Unification, *Economic Theory*, 18, 535-553.

[32] Nehring, K. (2002), Utilitarian Cooperation under Incomplete Information, mimeo UC Davis.*

[33] Nehring, K. (2003a), A Behavioral Characterization of Common Priors, mimeo UC Davis.*

[34] Nehring, K. (2003b),Utilitarianism without Common Priors, in preparation.

[35] Persico, N. (2003), Committee Design with Endogenous Information, *Review of Economic Studies,* forthcoming.

[36] Samet, D. (1998), Iterated Expectations and Common Priors, *Games and Economic Behavior* 24, 131-141.

[37] Segal, U. (2000), Let's Agree that all Dictatorships are Equally Bad, *Journal of Political Economy* 108, 569-589.

[38] Seidenfeld, T., J. B. Kadane and M. J. Schervish (1989), On the Shared Preferences of Two Bayesian Decision Makers, *Journal of Philosophy,* 86, 225-244.

[39] Wilson, R. (1978), Information, Efficiency and the Core of an Economy, *Econometrica* 46, 807-816.