Common Knowledge and Common Rationality Through Provability Logic[1]

Corrado Benassi[2] and Paolo Gentilini[3]

This Version: March, 1999

[1]Paper prepared for the GLS Conference on "Logic, Game Theory and Social Choice", Oisterwijk, The Netherlands, 13-16 May 1999.

[2]Dipartimento di Scienze Economiche, Università di Bologna, Piazza Scaravilli 2, 40141 Bologna, Italy.

[3]Irrsae Liguria, Via Lomellini 5, Genova, Italy

Abstract

The paper proposes a formalization of rational agents as first-order consistent formal systems. On this basis we build a notion of common knowledge and common rationality, among agents who are globally inconsistent with each other. An existence theorem for a formal system of common rationality is provided.

1 Introduction

In this paper we propose a logical formalization of the notions of knowledge, epistemic interaction and common knowledge among rational agents, the latter being formalized as first-order consistent formal systems. These notions allow us to introduce the idea of a system of common rationality for a society where agents have different and contrasting opinions (are globally inconsistent with each other in a formal sense), and are endowed with complex rationality (essentially, a proof-theoretic strength higher than that of recursive arithmetic PRA). By introducing some constraints on the agents' capability of proving the others' rationality (consistency), we are able to prove the existence of at least one formal system of common rationality (Theorem 1). We also show that some agents knowing that other agents are inconsistent, leads the former to negate the existence of a system of common rationality (Theorem2).

A crucial point of our approach is the idea of formalizing knowledge starting not from the modal system $S5$ (Geanakoplos, 1994), but from the modal system G, which is canonically isomorphic to the standard Provability Logic of Arithmetic. This enables us to use the provability predicate $\mathrm{Pr}_{T_j}(\cdot)$ of each system (agent) T_j as an epistemic operator, within the full setting of the proof-theory of the extensions of Arithmetics. There results a substantial refinement: since the provability of $\mathrm{Pr}_{T_j}(A)$ in general does not imply the Tj-provability of A, we can distinguish between a statement being in principle knowable, and its being actually known in a precise and detailed manner. In this sense, we believe that our framework might be a possible answer to the observation made by Parikh and Krasucki (1990), that one should look for formalizations of common knowledge which leave aside the knowledge of details (since such is life in "the real world").

The main tools we use are those of Proof Theory and Provability Theory; the upper bound to the rationality of agents is given in terms of proof-theoretic ordinals; however, many of the characterizations of knowledge we present are also based on Recursion Theory and Model Theory. Indeed, one of the main methodological arguments underlying our research programme is that theories of common knowledge and common rationality should take advantage of the resources provided by today's formal logic in a wide sense. Modal logic is of course among these resources (and we do work with the modal system G, canonically linked with Provability Logic), but we believe that the resources of proof theory - thus far essentially employed in the meta-analysis of mathematical systems - are much more powerful. In par-

ticular, proof theory provides us with infinite different provability predicates (which may represent different 'types' of individual rationality), and allows us to introduce a measure of rationality, in the form of measures of the logical complexity of formal systems. The latter we see as a crucial point, as different degrees of rational complexity allows to distinguish, on precise formal grounds, different epistemic processes.

We devote the rest of this introduction to introducing technical premises, notation and references. Section 2 concentrates on formalizing common knowledge through provability theory, Section 3 on developing a formal system of common rationality. Concluding remarks are gathered in Section 4.

1.1 Sequent formulation of the Predicate Calculus LK

We assume the reader to know the classical formulations and the basic tautologies of the First Order Predicate Calculus (Shoenfield, 1967; Mendelson, 1964). Since in our framework we identify the knowledge of a statement with its proof, and we want to express formally the knowledge-generating process, we shall use the Predicate Calculus in its Gentzen formulation as sequent calculus (Takeuti, 1987; Girard, 1987). Thus the proofs will be trees, whose leaves are axioms and whose branches are sequent rules.

We recall that a sequent is an expression of the form $X \Longrightarrow Y$, where X and Y are set of formulas. If $X = \{A_1, ..., A_n\}$, $Y = \{B_1, ..., B_m\}$, then $X \Longrightarrow Y$ has the same meaning of the formula $A_1 \wedge ... \wedge A_n \longrightarrow B_1 \vee ... \vee B_m$. The writing $\Longrightarrow A$ means that A is a theorem or an axiom; $A \Longrightarrow$means that $notA$ is a theorem or an axiom. Given a rule $\frac{S_1}{S_2}$, the sequent S_1 is the premise of the rule, the sequent S_2 is the conclusion of the rule. In a sequent we use Greek capitals as meta-expressions for sets of formulas, Latin for formulas.The writing Ω, Δ stands for $\Omega \cup \Delta$. The sequent formulation LK of the First Order Predicate Calculus is the following:

LK Axioms:

1. Logical Axioms: $A \Longrightarrow A$

2. Equality Axioms:

2.1 $s_1 = t_1, ..., s_n = t_n, A(s_1, ..., s_n) \Longrightarrow A(t_1, ..., t_n)$

2.2 $s_1 = t_1, ..., s_n = t_n \Longrightarrow f(s_1, ..., s_n) = f(t_1, ..., t_n)$

2.3 $\Longrightarrow s = s$

2.4 $s_1 = t_1, s_2 = t_2, s_1 = s_2 \Longrightarrow t_1 = t_2$

with s_j, t_j arbitrary terms, f function letter, A predicate letter.

3.Logical Rules:

3.1.Propositional logical rules:

$$\frac{A,\Gamma\Longrightarrow\Delta}{\Gamma\Longrightarrow\Delta,\sim A}\sim-\mathcal{R} \qquad \frac{\Gamma\Longrightarrow\Delta,A}{\sim A,\Gamma\Longrightarrow\Delta}\sim-\mathcal{L}$$

$$\frac{A,\Gamma\Longrightarrow\Delta}{A\wedge B,\Gamma\Longrightarrow\Delta}\wedge-\mathcal{L} \qquad \frac{B,\Gamma\Longrightarrow\Delta}{A\wedge B,\Gamma\Longrightarrow\Delta}\wedge-\mathcal{L} \qquad \frac{\Gamma\Longrightarrow\Delta,A \quad \Theta\Longrightarrow\Omega,B}{\Gamma,\Theta\Longrightarrow\Delta,\Omega,A\wedge B}\wedge-\mathcal{R}$$

$$\frac{\Gamma\Longrightarrow\Delta,A}{\Gamma\Longrightarrow\Delta,A\vee B}\vee-\mathcal{R} \qquad \frac{\Gamma\Longrightarrow\Delta,B}{\Gamma\Longrightarrow\Delta,A\vee B}\vee-\mathcal{R} \qquad \frac{A,\Gamma\Longrightarrow\Delta \quad B,\ \Theta\Longrightarrow\Omega}{A\vee B,\Gamma,\Theta\Longrightarrow\Delta,\Omega}\vee-\mathcal{L}$$

$$\frac{A,\Gamma\Longrightarrow\Delta,B}{\Gamma\Longrightarrow\Delta,A\longrightarrow B}\longrightarrow-\mathcal{R} \qquad \frac{\Gamma\Longrightarrow\Delta,A \quad B,\ \Theta\Longrightarrow\Omega}{A\longrightarrow B,\Gamma,\Theta\Longrightarrow\Delta,\Omega}\longrightarrow-\mathcal{L}$$

3.2. Quantifier logical rules

$$\frac{A(t),\Gamma\Longrightarrow\Delta}{\forall xA(x),\Gamma\Longrightarrow\Delta}\forall-\mathcal{L} \qquad \frac{\Gamma\Longrightarrow\Delta,A(b)}{\Gamma\Longrightarrow\Delta,\forall xA(x)}\forall-\mathcal{R}$$

$$\frac{A(b),\Gamma\Longrightarrow\Delta}{\exists xA(x),\Gamma\Longrightarrow\Delta}\exists-\mathcal{L} \qquad \frac{\Gamma\Longrightarrow\Delta,A(t)}{\Gamma\Longrightarrow\Delta,\exists xA(x)}\exists-\mathcal{R}$$

where in $\forall-\mathcal{L}$, $\exists-\mathcal{R}$, t is an arbitrary term, and in the corresponding $\forall xA(x)$, $\exists xA(x)$, t may occur, i.e. t may be not fully quantified. Conversely, in $\forall-\mathcal{R}$, $\exists-\mathcal{L}$, the free variable b occurring in A is uniformly replaced in $\forall xA(x)$, $\exists xA(x)$ by the bound variable x, and b does not occur in Γ,Δ; b is the *proper variable* of the rule.

4. Structural rules

4.1. Weakening rules

$$\frac{\Gamma\Longrightarrow\Delta}{\Gamma\Longrightarrow\Delta,A}w-\mathcal{R} \qquad \frac{\Gamma\Longrightarrow\Delta}{A,\Gamma\Longrightarrow\Delta}w-\mathcal{L}$$

4.2. Cut rule:

$$\frac{\Gamma\Longrightarrow\Delta,A \quad A,\ \Theta\Longrightarrow\Omega}{\Gamma,\Theta\Longrightarrow\Delta,\Omega} \quad Cut$$

In general we write $T=LK+A(T)+R(T)$ to indicate a system T obtained by adding to LK a proper axiom set $A(T)$ and a proper sequent rule set $R(T)$. The axioms B of $A(T)$ occur in the T-proofs as sequents of the form $\Longrightarrow B$.

It should be noticed that the form of the proof-trees in T describes the form and efficiency of agent T's reasoning.

1.2 Primitive Recursive Arithmetic $PRA(Z)$, Peano Arithmetic PA and Induction Rules

We suppose the reader is familiar with the notions of recursive function, recursive predicates, Turing machine and with basic Recursion Theory (Shoenfield, 1967; Odifreddi, 1989; Van Dalen, 1983).

Primitive Recursive Arithmetic PRA is the following first order formal system:

PRA language: the constant 0; a denumerable set of variables $x_1, ..., x_n, ...$; a denumerable set of function letters $f_1, ..., f_n, ...$, representing all recursive functions; a denumerable set of predicate letters $R_1, ..., R_n, ...$, representing all recursive relations.

PRA deduction apparatus:

LK

plus the following axioms (where t, t_i are arbitrary terms):

(a) $Z(t) = 0$ (Z zero function); (b) $\sim (S(t) = 0)$ (S successor function);

(c) $P_i^n(t_1, ..., t_n) = t_i$ (P_i^n projection function);

plus the following axiom schemata:

(d) composition axiom schema: given the functions $g_1, ..., g_n, h$ the axiom defines the function $f(x_1, ..., x_m) = h(g_1(x_1, ..., x_m), ..., g_n(x_1, ..., x_m))$;

(e) Primitive Recursion axiom schema: given the functions g and h the axiom defines the function f as follows: $f(x_1, ..., x_n, 0) = g(x_1, ..., x_n)$ and $f(x_1, ..., x_n, S(y)) = h(x_1, ..., x_n, y, f(x_1, ..., x_n, y))$;

plus the induction axiom schema or rule with no quantifier in the induction formula A. The induction axiom is $[A(0) \wedge \forall x(A(x) \longrightarrow A(S(x)))] \longrightarrow \forall x A(x)$. The sequent rule for induction we choose is

$$\frac{\Gamma \Longrightarrow \Delta, F(0) \qquad F(a),\ \Theta \Longrightarrow \Omega, F(S(a))}{\Gamma, \Theta \Longrightarrow \Delta, \Omega, F(t)} \quad \mathcal{I}$$

where, a is a free variable, called *proper variable* of the rule, not occurring in Γ, Δ, Θ, Ω, $F(t)$; F is an atomic formula; t is an arbitrary term which we say introduced by $\mathcal{I}$; $F(t)$ is the *principal formula* of $\mathcal{I}$; $F(0)$, $F(a)$, $F(S(a))$ are the *auxiliary formulas* of $\mathcal{I}$.

The system $PRA(Z)$ is a version of Recursive Arithmetic describing also the recursive operations and relations involving negative integers; we introduce a unary function letter $p(\cdot)$, the *predecessor function* (with the

intended meaning $p(m) = m - 1$), and axiomatize both successor and predecessor functions by the following axioms, where $t_1$and t_2 are arbitrary terms: (i) $S(p(t)) = t$; (ii) $p(S(t)) = t$; (iii) $S(t) = 0 \longrightarrow t = p(0)$; (iv) $p(t) = 0 \longrightarrow t = S(0)$; (v) $S(t_1) = S(t_2) \longrightarrow t_1 = t_2$; (vi) $p(t_1) = p(t_2) \longrightarrow t_1 = t_2$. The functions *zero* $Z(\cdot)$, projection P_j^k, an extended recursion schema and the composition schema can all be added straightforwardly. When we work in $PRA(Z)$, we refer to the following three-premise sequent version of the induction rule:

$$\frac{\Gamma \Longrightarrow \Delta, F(0) \qquad F(a),\ \Theta \Longrightarrow \Omega, F(p(a)) \qquad F(a),\ \Theta \Longrightarrow \Omega, F(S(a))}{\Gamma, \Theta \Longrightarrow \Delta, \Omega, F(t)} \quad \mathcal{I}$$

where the constraints established for the PRA Induction Rule extend to the $PRA(Z)$ Induction Rule straightforwardly.

The system of *Peano Arithmetic* PA is here defined as the system $PRA(Z)$ extended by induction axioms or induction rules, *where in the induction formula an arbitrary number of quantifiers may occur.* A central remark is the following: *formal induction yields a substantial qualitative increase in the proof-theoretic strength of a formal system.* In particular, assigning an induction rule to a theory implies endowing it with a sort of meta-inference capability, which allows it to prove the consistency of some of its subsystems. PRA can prove the consistency of Arithmetic without induction, and PA can prove the consistency of PRA, induction being essential in these proofs. Moreover: the proof-theoretic strength of a PA-subsystem T can be graded by the highest number of quantifier occurrences allowed in the induction formulas in the T-proofs (see Takeuti, 1987, p.116). We indicate as $PRA(Z)_k$ the system obtained by extending $PRA(Z)$ with induction rules or axioms where k quantifiers at most occur in the induction formula.

The problem of the consistency of PA is not simple; as we shall see in 1.3 below, the proof of the consistency of PA and its subsystem $PRA(Z)_k$, $k > 0$, can be syntactically proved by transfinite induction on ordinals greater than ω (and anyway *via* procedures not strictly finitistic, but much more constructive than semantic methods).

Also, we introduce in PA the following technical extension of the first order language: following Troelstra and Schwichtenberg (1996, p.263), we add a denumerable set $C_1, ..., C_n, ...$ of second-order unary formula variables which act as free set variables. *We assume that these are never quantified in any proof of our systems.* Through these second order variables the rule schemata and the axiom schemata can be considered as included in the

PA language. Hence, since we work with systems $T_j, j = 1, ..., m$ which are recursively axiomatized, the notions "rule schema in the system T_j", "axiom schema in the system T_j" can be expressed by recursive predicates. We shall for short speak of "T_j-rule" or "T_j-axiom", whenever the context makes it clear whether we mean a schema (where formula variables occur), or the instance of a schema (where only first-order terms and predicate constants occur). As to model theory, since this language extension yields no relevant extension with respect to the first-order theorems of PA, we consider only the models of the first-order theorems of the PA-oversystems we work with.

1.3 Transfinite Induction up to ordinals below ε_0 and the proof-theoretic strength of a system

Consider a set A endowed with a binary relation $\prec$ which is irreflexive, transitive and linear (i.e., for any $x, y \in A$ either one of the following must hold: $x \prec y$, $y \prec x$, $y = x$). This is a *well ordering* if each non-empty subset X of A has a $\prec$-minimum. *Ordinals* are sets well ordered by the set-theoretic binary relation $\in$, and each ordinal represents an equivalence class, by isomorphism, of well orderings. Natural numbers represent all the finite ordinals. The first infinite ordinal is N, also indicated by ω; beyond ω we have the class of infinite ordinals, usually denoted with Greek letters like η, ξ, ... The schema of transfinite induction on ordinals up to a fixed β corresponds to the following inference: let $F(\gamma)$ be a statement depending on ordinals; *if* for each $\alpha \prec \beta$, $F(\xi) \longrightarrow F(\alpha)$ for each $\xi \prec \alpha$, *then* $F(\alpha)$ holds for each $\alpha \prec \beta$ (see, also for the definitions of ordinal sum and ordinal exponentiation: Takeuti, 1987; Girard, 1987; Pohlers, 1989; Troelstra-Schwichtenberg, 1996).

The consistency of PA is provable by transfinite induction up to the countable ordinal ε_0, which is so defined: *if* $\omega(0) \equiv 1$ and $\omega(n+1) \equiv \omega^{\omega(n)}$, *then* $\varepsilon_0 \equiv \sup_n \{\omega(n)\}$. It is possible to formalize ordinals below ε_0 inside the PA language, by means of a canonical bijection Φ between the set $\{\alpha : \alpha \prec \varepsilon_0\}$ and the non-negative integer set N (Troelstra-Schwichtenberg, 1996, p.262); Φ allows to define in N *a well ordering of order type* ε_0, which reproduces in N the well ordering of ε_0. We denote by $\angle$ such ε_0-well ordering in N, and we introduce $\angle$ in the PA language. Moreover, following Troelstra and Schwichtenberg (1996, p.264) we add to PA a set of axioms we denote *Ord* describing the properties of $\angle$, and hence of the PA-translation of ordinals below ε_0. For simplicity, we use the Greek letters η, ξ, ...to indicate also the elements of N representing infinite ordinals according to

the bijection Φ, but we should recall that they are in fact the numerals $\Phi(\eta)$, $\Phi(\xi)$,...of the PA language. Given all this, the axioms of transfinite induction up to ε_0 in the PA language is

$$\forall x\left(\forall y \angle x F(y) \longrightarrow F(x)\right) \longrightarrow \forall x F(x)$$

the axiom of transfinite induction up to α, $\omega \angle \alpha \angle \varepsilon_0$, in the PA language is

$$\forall x\left(\forall y \angle x F(y) \longrightarrow F(x)\right) \longrightarrow \forall x \angle \alpha F(x)$$

The corresponding sequent rules are the following

$$\frac{a \angle b, F(a), \Gamma \Longrightarrow \Delta, F(b)}{\Gamma \Longrightarrow \Delta, F(t)} \qquad \frac{\Gamma \Longrightarrow \Delta, A}{t \angle \alpha, \Gamma \Longrightarrow \Delta, F(t)}$$

where a and b are free variables which do not occur in Γ, Δ, $F(t)$, and t is an arbitrary term. We indicate with $I(\beta)$ the transfinite induction axiom up to the ordinal β, $\omega \angle \beta$, and with $IR(\beta)$ the corresponding sequent rule. We have that $PRA(Z) + Ord + I(\varepsilon_0)$ proves the consistency of PA, and that $PRA(Z) + Ord + I(\omega(k+1))$ is sufficient to prove the consistency of $PRA(Z)_k$(Takeuti, 1987, p.116).

A theory T which is a $PRA(Z)$-extension can be characterized by the least ordinal which is necessary to prove its consistency by transfinite induction. We call such ordinal the *measure of the complexity of theory* T. The complexity of $PRA(Z)$ is ω, the complexity of PA is ε_0. We call *ordinal measure of the proof-theoretic strength* of a consistent PA-extension V, the maximum ordinal of a transfinite induction rule which V can derive. E.g., $PRA(Z) + Ord + I(\omega(k+1))$ has proof-theoretic strength of measure $\omega(k+1)$.

Let $P_1, ..., P_m$ be formal proofs in the PA-extensions $T_1, ..., T_m$, where the sets of induction rules (both standard and transfinite) $\mathcal{J}_1, ..., \mathcal{J}_m$ occur; we say that the sets $\mathcal{J}_1, ..., \mathcal{J}_m$ have the same strength if the system $PRA(Z) + Ord + \mathcal{J}_r$ has the same theorems for each $r = 1, ..., m$.

We conclude this section by noting that, if we think of a system V as an epistemic agent: (a) its ordinal and the ordinal measure of its proof-theoretic strength may be read as a measure of the complexity of V's knowledge and his reasoning powers; (b) we consider transfinite induction up to ε_0 as the formal expression of the limit of the knowledge V can effectively reach.

1.4 Gödel's theorems, basic Provability Logic, and the modal system G

We recall that by Gödel numbering we can injectively assign a natural number $\ulcorner E \urcorner$ to each expression E of the language of a formal system. The *provability predicate* for a fixed recursively axiomatized first-order theory T is the PRA formula $\exists x Prov_T(x, \ulcorner E \urcorner)$, meaning "there exist in T a proof of the formula coded by the number $\ulcorner E \urcorner$", for which we introduce a new predicate letter $\mathrm{Pr}_T(\cdot)$: we write for short $\mathrm{Pr}_T(E)$, omitting $\ulcorner \cdot \urcorner$. Note that $\mathrm{Pr}_T(\cdot)$ is in general recursively enumerable, but not recursive. Moreover, it is indexed by the theory T: changing T, its properties change as well. Nevertheless, for the class of systems T which are PRA-extensions, $\mathrm{Pr}_T(\cdot)$ has important standard properties indipendent of T. Hence, PRA can describe (though not necessarily prove) consistency of all recursively axiomatizable theories T. Indeed, if the symbol $\bot$,'falsum', is included in the T-language (by defining $\bot \longleftrightarrow A \wedge \sim A$), we have that $\sim \mathrm{Pr}_T(\bot)$ expresses in PRA the consistency of T, while $\mathrm{Pr}_T(\bot)$ expresses in PRA the inconsistency of T. We shorten as $Coer(T)$ the formula $\sim \mathrm{Pr}_T(\bot)$.

The well known Gödel's incompleteness theorems are the following:

I. Assuming PA-consistency, there is a PA-sentence A such that neither $\vdash_{PA} A$, nor $\vdash_{PA} \sim A$.

II. For each consistent recursively axiomatized extension T of PRA, $\nvdash_T Coer(T)$, i.e. T cannot prove its own consistency.

Provability Logic (Solovay, 1976; Smorynski, 1977 and 1985; Boolos, 1979; Gentilini, 1992, 1998) is the logic expressing the properties of the provability predicate $\mathrm{Pr}_T(\cdot)$, when T is a recursively axiomatized extension of PRA. A key fact is that such properties are T-provable: as a consequence, though T cannot prove its own consistency, it can prove a wide set of significant statements involving its consistency and the notion of T-provability. The following are standard provability logic statements:

1. $\vdash_T A$ implies $\vdash_T \mathrm{Pr}_T(A)$;
2. $\vdash_T \mathrm{Pr}_T(A \longrightarrow B) \longrightarrow (\mathrm{Pr}_T(A) \longrightarrow \mathrm{Pr}_T(B))$;
3. $\vdash_T \mathrm{Pr}_T(A) \longrightarrow \mathrm{Pr}_T(\mathrm{Pr}_T(A))$;
4. $\vdash_T \mathrm{Pr}_T(A) \longrightarrow A$ iff $\vdash_T A$,
   formalized as $\vdash_T \mathrm{Pr}_T(\mathrm{Pr}_T(A) \longrightarrow A) \longrightarrow \mathrm{Pr}_T(A)$;
5. $\vdash_T \mathrm{Pr}_T(\sim \mathrm{Pr}_T(A)) \longleftrightarrow \mathrm{Pr}_T(\bot)$.

Note that a weaker theory can prove provability logic statements concerning a stronger theory: e.g., $\vdash_{PRA} \mathrm{Pr}_{PA}(B)$, for each PA-theorem B. Note also that the provability predicate can be seen as an epistemic operator: $\mathrm{Pr}_T(A)$ can be read as "agent T thinks that A"; in this vein, $Coer(T)$ expresses in T what is unknowable by T.

The system of propositional modal logic G (Boolos, 1979; Smorynski, 1985; Gentilini, 1992, 1998) is based on the following axioms and rules: (GA.1) *all propositional tautologies*; (GA.2) $\Box(A \longrightarrow B) \longrightarrow (\Box A \longrightarrow \Box B)$; (GA.3) $\Box(\Box A \longrightarrow A) \longrightarrow \Box A$; (GR.1) from $A, A \longrightarrow B$ infer B (modus ponens); (GR.2) from A infer $\Box A$ (necessitation).

We know that G can be formulated in a sequent formalism, such as Propositional Calculus plus the rule

$$\frac{\Gamma, \Box\Gamma, \Box A \Longrightarrow A}{\Box\Gamma \Longrightarrow \Box A}$$

The system G is the Provability Logic modal system, since if we define the interpretation of a modal language $\mathfrak{L}$ as an application $\phi : \{\mathfrak{L}\text{-}formulas\} \longmapsto \{PA\text{-}formulas\}$ which preserves propositional connectives and such that $\phi(\Box A) = \mathrm{Pr}_{PA}(\phi(A))$, we have the following completeness theorem (Solovay, 1996; Gentilini, 1998):

$$\vdash_{PA} \phi(A) \text{ for each } \phi \text{ implies } \vdash_G A$$

The provablity interpretation of the G-theorem $\Box(\Box A \longrightarrow A) \longleftrightarrow \Box A$ means that the provability of A (A arbitrary) is equivalent to the provability of $\Box A \longrightarrow A$. This implies that $\Box A \longrightarrow A$ cannot be a G-theorem. For this reason, if $\Box$ is seen as an epistemic operator, A being knowable does not imply A.

1.5 The standard model $\mathcal{Z}$ and the non-denumerable infinity of non standard models of Arithmetic

We assume the reader is familiar with the basic Tarskian semantic for first order logic (Shoenfield, 1967; Chang-Kiesler, 1973; Barwise, 1977). Recall that $PRA(Z)$ e PA admit a non-countable infinity of non isomorphic models, among which lies the standard model $\mathcal{Z}$, made of by the set Z of integers with the usual arithmetic operations and relations. From our standpoint, it is important to stress that there are nonstandard models expressing truth notions very different from the usual one: e.g., there are

models where $\mathrm{Pr}_{PA}(\perp)$ is true, and $Coer(PA)$ is false. Thus, the system $PRA(Z)+Coer(PRA(Z)_k)+\mathrm{Pr}_{PA}(\perp)$ is consistent, has an infinity of non isomorphic models, and *does not preserve the standard model*, in the sense that there is no expansion of $\mathcal{Z}$ among its models (an expansion $\mathcal{A}'$ of the structure $\mathcal{A}$ for a language L is a structure for an over-language $L' \supset L$, which preserves the interpretation of the symbols of L).

2 Expressing knowledge through first order systems

We assume that the knowledge of a rational agent can exist only as a set of expressions in a developed natural language. The latter is formalized as the formal language of first-order logical systems, which is assumed to include the language of $PRA(Z)$: we think of the possibility of numbering, performing numerical exercises, recognizing recursive relations among objects and quantities, as necessary components of rationality. Hence, we see a state of the world σ as a set of first-order sentences, which are knowledge only iff they are part of a rational agent as axioms or theorems of his.

The environment within which knowledge takes place (both at the individual level, and as common knowledge when epistemic interaction is allowed for), is a finite society of rational agents T_j, each of which is a first-order logical system. Differences in the logical complexities and proof-theoretic strengths of the different T_j's will define (possibly very largely) different states of knowledge and rationality of society.

For now, we limit ourselves to defining the minimal properties with which each T_j is endowed, such that individual knowledge within society can be represented:

Definition 1 *Each agent T_j of society $S \equiv \{T_1, ..., T_m\}$ is a first-order logical system such that:*
(a) the language L of T_j is common to all j's and includes the language of $PRA(Z)$;
(b) for all $T_j \in S$, both the set $A(T_j)$ of the T_j-axioms, and the set $R(T_j)$ of T_j-inference rules are non empty, and such that T_j is consistent and proves at least all the theorems of the first order predicate calculus LK, which can be exressed in the language L;
(c) there exists in S no pair of agents who are equal: the proper axiom set $A(T_j)$ and the proper inference-rule set $R(T_j)$ are different for any pair

(T_j, T_k), $j \neq k$; this implies that for any T_j there exists an infinite denumerable set of T_j-theorems which are not T_k-theorems for any $k \neq j$.

This definition imposes that each T_j be endowed with a specific rationality, and hence a set of specific knowledge, different from those of other agents. However, the constraint of the common language L and on the common part of the deductive apparatus represented by LK, allows that different items of knowledge can be communicated and compared. This is not really common knowledge as will be defined later, but it is a pre-requisite of it. Our notion of individual knowledge can be formalized as follows.

Definition 2 *Given society $S \equiv \{T_1, \ldots, T_m\}$, we say that $T_j \in S$ knows the part of the universe Ω represented by sentence A, iff $\vdash_{T_j} A$, i.e. iff T_j proves A using his deductive apparatus $A(T_j) + R(T_j)$.*

Notice that the common basis on the states of the world which can be considered as given or starting points by T_j, can be assumed to be included in the set of his proper axioms $A(T_j)$; on the other hand, axioms are themselves theorems, and our definition of knowledge is inferential and theoretical in character: even the sheer observation of a given state of the world amounts to subsuming that observations as an axiom.

This definition of knowledge allows us to distinguish formally among states of individual knowledge which are very different. A key role in such a distinction is played by the provability predicate $\mathrm{Pr}_{T_j}(\cdot)$, which (as we have seen) is canonically linked to the modal system G. The Provability Logic of the predicate $\mathrm{Pr}_{T_j}(\cdot)$ includes the peculiarities of T_j's deductive (and cognitive) apparatus, and can express self-reference of knowledge, i.e. knowledge of one's (and the others') knowledge. This is so, because its properties are finer than those of the kripkian epistemic operator K, linked to the modal system $S5$ (Geanakoplos, 1994). Indeed, the system G can formalize the difference between T_j knowing A (i.e., $\vdash_{T_j} A$), and T_j's knowledge that A can be known by T_k, $k \neq j$ (i.e., $\vdash_{T_j} \mathrm{Pr}_{T_k}(A)$).

If one were to define a universe Ω of all that can be known, this would certainly include (like in the standard approach of the common knowledge literature) physical objects and their properties, the very knowledge of rational agents, the agents' reasoning about their own and the others' knowledge, actions by agents, knowledge of one's and the others'actions, and so on. However, we surmise that these are in principle different levels of knowledge, which are characterized by the formal structure of the sentence A in

the "act of knowing" we formalized as $\vdash_{T_j} A$, as well as by the proof-strength which is necessary in T_j to prove A. We can at this point summarize the different meta-levels of individual knowledge which can be formalized in our framework:

1. *Self-evident knowledge* can be identified when A is a *recursive* relation (Van Dalen, 1983; Odifreddi, 1989), expressing effective relations between objects which can be codified by numbers. If all $T_j \in S$ include the $PRA\,(Z)$, we have $\vdash_{T_j} A$, implies $\vdash_{T_i} A$ for all $i = 1, \ldots, m$: hence, all agents have the same knowledge about computable relations between objects.

2. Suppose we codify a state of the world σ (a set of first order sentences) by a vector $s \in Z^r$, with finite r. An action of agent T_j is represented by a function $f_j : Z^r \longmapsto Z^r$,expressing the change in the world induced by the action. T_k's knowledge of an action by T_j will be given by $\vdash_{T_k} A$, with $A \equiv (f_j\,(s) = s')$, that is the value of the nonrecursive action function f_j axiomatized by T_j.

The function f_j being non recursive establishes that it is *individual* knowledge and behaviour, in that there is no effective way a recursive apparatus external to T_j can compute it. This requires that T_j be more complex than a Turing machine, and that it be an undecidable logical system (Odifreddi, 1989; Shoenfield, 1967). Both these features are implied by Definition 1. Undecidibility has an important epistemic meaning: it says that there are no effective procedure (external to the agent) which can establish whether a sentence is or is not within his individual knowledge. Moreover, the action function f_j will be axiomatized by the most developed part of the set of proper axioms $A(T_j)$: hence, for $k \neq j$, T_k will not generally be able to prove T_j's actions: he cannot know them explicitly, although (given the common language) T_k will be able to prove theorems which include the functional letter f_j as an object (i.e., to speculate about T_j 's actions and have a subjective knowledge of them).

3. T_j's awareness that T_k knows something expressed by B is represented by

$$\vdash_{T_i} \mathrm{Pr}_{T_k}(B)$$

that is, T_j proves that T_k proves B: the former proves (knows) that B can be proved (known) by the latter. Notice however that this does not imply at all that T_k actually knows B, *since for any* A, $\mathrm{Pr}_T(A) \longrightarrow A$ *is a provability logic theorem of no system* T. Indeed, in general it will not be the case either,

that $(\vdash_{T_j} \mathrm{Pr}_{T_k}(A)) \longrightarrow A$ for $j \neq k$, because whenever A is inconsistent with T_j and is a theorem of T_k, this would deliver inconsistency of T_j. Accordingly, the two following formulas may in general hold simoultaneously

$$\vdash_{T_j} \mathrm{Pr}_{T_k}(B) \, and \nvdash_{T_j} B, \quad j \neq k$$

Thus the provability predicate allows us to distinguish between a statement of T_j concerning whether B can be known by T_k, and T_j's direct knowledge of B. On this point we make two observations:

(a) This distinction lies at the heart of our definitons of common knowledge and common rationality. Thus, if $f_k : Z^r \longmapsto Z^r$ is T_k's action function, T_j's awareness of T_k's action will be given by

$$\vdash_{T_j} \mathrm{Pr}_{T_k}\left(f_k(s) = s'\right), \; j \neq k$$

where T_j is not required to know f_k precisely or explicitly. The latter we see as a crucial point: in social intercourse, the more developed is individual knowledge (to the point of having the others' knwoledge and actions as its objects, as is the case with economic or political choices), *the more common knowledge is characterized by a drastic reduction of data and details, and the less is characterized by actual sharing of knowledge* (which is arguably not the case with animal cognitive processes). A case can be made, that in human society the building of a common rationality is based on individual rationality being able to leave aside the details of one's individual knowledge, and speculate whether something can be known without actually knowing it. Suppose for example that T_k is a candidate to the US presidency, and T_j someone who could vote for him; and suppose that B is a detailed description of the foreign policy position T_k would advocate, once elected, on some particular issue. Clearly, T_j is neither interested in, nor able to acquire, a precise and detailed knowledge of B; however, he would require knowledge of $\mathrm{Pr}_{T_k}(B)$ – a US president should be prepared on a foreign policy issue. Indeed, this will hold for almost all points of T_k's platform: in practice T_j will support that platform only if he gives up knowing all the details thereof. On the other hand, it seems legitimate to claim that *T_k's electoral platform is commonly rational, which implies some form of common knowledge among all electors (at least those who vote for him) in society S.* This will be formalized in section 3, but in this section we wanted to outline the crucial

role of the provability predicate Pr as an epistemic operator different from K of $S5$, which may allow to define interesting notions of common rationality.

(b) Metalevels of knowledge about the others' knowledge amount to conjectures. Indeed, the two propositions $\vdash_{T_j} \Pr_{T_k}(B)$ and $\nvdash_{T_j} B$ $(j \neq k)$ can co-exist without T_j's consistency being jeopardized (to be sure, one can also have $\vdash_{T_j} \backsim \Pr_{T_k}(B)$ and $\vdash_{T_j} B$). In this case Pr is used as a credence operator. E.g., let B be Goldbach's conjecture: the sentence $\vdash_{T_j} \Pr_{T_1}(\Pr_{T_4}(\Pr_{T_3}(B)))$ can be read as T_j stating that T_1thinks that T_4 thinks that T_3 has proved Goldbach's conjecture. Obviously, this can be subjectively 'known' by T_j, even though T_1 and T_4 do not actually think what is being said they think.

4. Agent T_j 's reasoning about knowledge and actions (his own and the others') is represented by complex sentences, where predicates like $\Pr_{T_k}$ and $\Pr_{T_j}$ will occur ($j, k = 1, \ldots, m$), and by the sequence of theorems by which such sentences are proved within T_j.

3 Common Knowledge as Epistemic Interaction

Within a society of complex rational agents, the notion of common knowledge will require the existence of individual 'knowledges' which are not only different, but also logically incompatible with each other: we should allow for different opinion about the world to exist. We shall accordingly concentrate on a definition of common knowledge and common rationality within a *democratic society*:

Definition 3 *The society* $S \equiv \{T_1, \ldots, T_m\}$*is democratic iff for any pair of agents* (T_j, T_k), $j \neq k$, *there exists a sentence* A *such that* $\vdash_{T_j} A$ *and* $\vdash_{T_k} \sim A$ *or, equivalently by the Craig-Robinson theorem (Shoenfield, 1967), iff the system* $T_j \cup T_k$ *is inconsistent*

This amounts to any individual's knowledge being globally incompatible with any other's - which indeed is what makes it interesting to define common knowledge. However, if we want the logical systems T_j to represent complex agents in such a society, we shall have to be precise about the upper and lower bounds for their proof-theoretic strengths and expressive capabilities. Accordingly, we define the following level of inductive rationality for an agent:

Definition 4 *We say that the logical complexity of agent $T_j \in S$ is at the level of inductive rationality iff the following conditions are satisified:*
(a) The language $L(T_j)$ is a proper extension of the language of $PRA(Z)$, with non recursive function letters $f_1, \ldots, f_m$ expressing an agent's action function.
(b) For all $j = 1, ..., m$, $A(T_j)$ and $R(T_j)$ ensure that T_j be consistent and recursively axiomatizable (Shoenfield, 1967; Smorynski, 1977); moreover, T_j includes the sequent version of $PRA(Z)$ and the axioms Ord for the basic properties of ordinals less than ε_0.
(c) Each T_j is endowed with either an axiom schema of classical induction over N, or a sequent-formulated classical induction rule over N, possibly with a bound r on the number of quantifiers occurring in the induction formula. This implies T_j proves all theories of standard provability logic of each recursively axiomatized system.
(d) Each T_j has either an axiom schema $I(\beta)$, or a sequent-formulated rule $IR(\beta)$, of transfinite induction up to a denumerable ordinal β different from ω, $\omega \angle \beta$, possibly with a bound r on the number of quantifiers occurring in the induction formula; β is characterized by the following proof-capabilities of the resulting system T_j:

(i) T_j proves the consistency of at least a system $PRA(Z)_k$, strictly included between $PRA(Z)$ and PA;

(ii) T_j does not prove the consistency of PA, which implies $\beta \angle \varepsilon_0$.
(e) The rules $R(T_j)$ are sound in the following sense: if M is any model of the axioms $A(T_j)$, and any rule premise is M-true, then the rule conclusion is also M-true.

Some comments about this definition are as follows:

1. Reasoning is formalized as a proof-tree, whose leaves are axioms and whose root is the proven theorem. Within T_j's inference, the distribution between axioms and rules expresses the form of T_j's reasoning, and bears obviously on the length of knowledge-processing; we assume that the basic (LK-predicate) inferences are anyway carried out *via* the rule of sequent calculus. A preference for induction axioms over induction rules requires a heavier use of the cut rule, and hence introducing in the proof complex formulas which have to be cut over the root - apparently, a less efficient approach.

2. The system T_j is not required to preserve an expansion of the standard model $\mathcal{Z}$ of $PRA(Z)$ among its models. In general, we allow non standard

"rationality", provided consistency is assured: e.g., an agent can prove for some k a result like $\mathrm{Pr}_{T_k}(\bot)$, stating T_k's inconsistency, even though this is false in $\mathcal{Z}$. In this sense, an agent's *subjective* knowledge (or awareness) may differ from the possible notion of objective "truth" formalized by the standard model $\mathcal{Z}$.

3. As a straightforward corollary of our definition is that, if T_j is at the level of inductive rationality, then it is necessarily undecidable and syntactically incomplete: *a fortiori*, Gödel's theorems on the non provability or refutability of its own consistency hold for T_j.

4. Finally, there exists a basic infinite set of theorems *all* agents prove (things they know), since each T_j includes $PRA(Z)$. However, S being democratic (Def. 3), T_j *cannot know the whole of* T_k*'s individual knowledge* (Def.2), $k \neq j$. Indeed, T_j is in general inconsistent with it: there is a denumerable infinity of statements A such that if $\vdash_{T_k} A$, $\nvdash_{T_j} A$ holds - things known by T_k but unkowable by T_j, as is the case when A is an action by T_k. However, the following is an essential, if straighforward, consequence of the provability logic included in all agents: *if* T_k *knows* B*, then any other agent* T_j *knows that* B *is* T_k*-knowable*, even if T_j does not know B - i.e., $\vdash_{T_j} \mathrm{Pr}_{T_k}(B)$ holds, even if in general $\nvdash_{T_j} B$. As a result, there are events that every agent is aware can be known by society S, even though he knows directly only a small part of them. Though arguably paradoxical, this is a clear upshot of the mathematical formalism. Notice however that knowing that other know is not enough to characterize common knowledge: as we shall see, the set $\{\mathrm{Pr}_{T_k}(B) : \vdash_{T_k} B,\ k = 1, ..., m\}$ *is not* equivalent to the common knowledge of society $\{T_1, ..., T_m\}$, which we interpret as arising from epistemic interaction.

We shall confine our treatment of common knowledge to democratic societies whose members are inductively rational. The common knowledge relevant in this setting is borne out by interaction among agents: accordingly, we do not give much weight to the set of common theorems:

Definition 5 *Given a democratic society* S *whose agents are at the level of inductive rationality,* trivial *common knowledge is the set of sentences* $\cap_{j=1}^{m} \{T_j\text{-}theorems\}$

Lacking explicit mention to the contrary, in the sequel we shall mean by "common knowledge" non-trivial common knowledge. In our framework, the

latter is the acquisition of others' opinions (different from one's opinions), without losing one's consistency. Thus:

Definition 6 *Let S be a democratic society with agents at the level of inductive rationality. Then a sentence A is* potential common knowledge *if the following conditions hold:*
(a) A is a theorem of at least one T_k , but $A \notin \cap_{j=1}^{m} \{T_j\text{-theorems}\}$, so that $\nvdash_{T_j} A$, for some $j \neq k$. Hence, $\vdash_{T_j} \mathrm{Pr}_{T_k}(A)$ for each $j = 1, ..., m$ and each k such that $\vdash_{T_k} A$;
(b) A preserves the standard model $\mathcal{Z}$ of the basic system $PRA(Z)$ included in all T_j's, in the following sense: $PRA(Z) + A$ is consistent and has an expansion of $\mathcal{Z}$ among its models.

As one can see, point (b) of Definition 6 introduces a condition which is semantic, based on the formal models of T_j. So we notice that in a democratic society there is no model which is common to all agents – indeed, no pair (T_i, T_j) admits of a common model, otherwise $T_i \cup T_j$ would be inconsistent. Moreover, we already observed that every T_j (as an extension of $PRA(Z)$) admits of an infinity of non-isomorphic models, among which the standard model $\mathcal{Z}$ is not necessarily preserved. E.g., the system $PRA(Z) + \mathrm{Pr}_{PRA}(\bot)$ is consistent and admits of infinite nonisomorphic models, but among the latter the standard model given by the set Z of integers and their ususal operations is not to be found – indeed, in this model the formula $\mathrm{Pr}_{PRA}(\bot)$ cannot be true, since it states the inconsistency of PRA, a system for which we do have a constructive proof of consistency. Thus, point (b) is a naturality requirement for a sentence which is a candidate for common knowledge – that is, a requirement of truth in the natural model of the basic deductive system. As is well know, $PRA(Z) + A$ being consistent does not imply that $T_j + A$ be consistent (in the case where $\nvdash_{T_j} A$); neither does $PRA(Z) + A$ preserving the standard model of $PRA(Z)$ imply that the same is preserved by the $PRA(Z)$-extension T_j, even if A is a T_j-theorem. At the same time, recall that we do not constrain our agents to preserve the standard model $\mathcal{Z}$ (i.e., we do not limit individual rationality in this sense), but we ask that for common knowledge the truth be holding with respect to the standard model. Hence, T_k can prove theorems like $\mathrm{Pr}_{T_j}(\bot)$, $j \neq k$, but these do not enter common knowledge, even if they are true in the infinte models of T_k.

The following definition provides a useful technical benchmark:

Definition 7 *The sentence B of the language of $S \equiv \{T_1, ..., T_m\}$ is* complex, *iff it is neither a $PRA(Z)$-theorem, nor a negation of a $PRA(Z)$-theorem, and moreover a finite $k \geq 1$ exists, such that $\vdash_{PRA(Z)+Ord} B \longrightarrow Coer\,(PRA\,(Z)_k)$*

The complexity lies in that, if $\vdash_{T_j} B$, then T_j proves $Coer(PRA\,(Z)_k)$, since $PRA(Z) + Ord$ is included in every T_j.

We are now in the position to define epistemic interaction and actual common knowledge as follows:

Definition 8 *Let $S \equiv \{T_1, ..., T_m\}$ be a democratic society with agents at the level of inductive rationality. Let $\mathcal{B} \equiv \{B_1, ..., B_m\}$ be a m-ple of sentences such that: (a) each B_j is potential common knowledge for S; (b) $\vdash_{T_j} B_j$ and $\nvdash_{T_j} B_k$, for each $j, k = 1, ..., m$, $k \neq j$; (c) $\mathcal{B}$ admits a common model M which is an expansion of the standard model $\mathcal{Z}$ of $PRA(Z)$, which implies the consistency of the system$PRA(Z) + \wedge_{r=1,...,m} B_r$.*
An epistemic interaction in S is a m-ple $(T_1 + \{B_r\}_{r\neq 1}, ..., T_m + \{B_r\}_{r\neq m})$ such that the following homogeneity condition *is satisfied:*
either no complex sentence is in $\mathcal{B}$;
*or $\mathcal{B}$ includes only complex sentences: then a maximum k exists such that $\vdash_{PRA(Z)+Ord} B_r \longrightarrow Coer\,(PRA\,(Z)_k)$ for each r, and the set $\mathcal{J}_r$ of induction rules used by T_r to prove B_r has the same strength for every r(see*1.3*), i.e.$PRA(Z) + Ord + \mathcal{J}_r$ has the same theorems for all r; in this latter case epistemic interaction is said to be* complex.
An epistemic interaction is an agreement, if all $T_1 + \{B_r\}_{r\neq 1}, ..., T_m + \{B_r\}_{r\neq m}$ are consistent; a disagreement otherwise.
If the epistemic interaction is an agreement, then every set $\{B_r\}_{r\neq k}$ is actual common knowledge *in S in the perspective of T_k.*

Remark 1 *Property (c) above does not in itself imply that the set $\mathcal{B}$ be consistent with some T_j.*

According to our definition, then, in this framework actual common knowledge is defined in the perspective of T_k. On the one hand, T_k is aware that the commonly known statements can be known by others agent; on the other hand, the homogeneous epistemic interaction is such that all commonly known statements are simultaneously true in the same expansion of the standard model, and at the same time consistent with each agent. This (perhaps) restrictive definition delivers a notion of common knowledge, according to which each T_k knows that others know, without the constraint of knowing precisely what they know.

4 Common rationality and its existence

We now want to take up formally the intuitive notion of "common way of thinking" in a group of rational agents within a democratic society. Actual common knowledge as defined in the former section is a set of sentences, and not a deductive apparatus. Let us go back to the example of the voters $\{T_1, \ldots, T_m\}$ and the candidate for the US presidency in Section 2. We can certainly say that the set of sentences $\mathcal{B}$ describing the candidate's electoral platform is potential common knowledge as defined in Definition 6; moreover, it generates epistemic interaction and agreement according to Definition 8. Thus we can say that some relevant common knowledge is produced by the platform, even though each voter has a detailed knowledge only of a small part of it. We have not, as yet, formalized the idea that such a platform may represent a common way of reasoning on some part of the (social) universe, but not on the whole of it.

Our definition is as follows:

Definition 9 *Let $S \equiv \{T_1, ..., T_m\}$ be a democratic society with agents at the level of inductive rationality; a sistem U of* common rationality *for S is a first-order logical system with the following properties:*
(a) U is consistent, and each system $T_j + U$ is consistent;
(b) For each T_j, at least one proper axiom in $A(T_j)$ which is not a theorem of T_k, $k \neq j$, and at least one inference rule in $R(T_j)$ or T_j-derivable which is not in the basic system $PRA(Z) + Ord$, are respectively in the proper axiom set $A(U)$ and in the rule set $R(U)$. Nothing else is part of $A(U)$ and $R(U)$; U includes strictly $PRA(Z) + Ord$ and is recursively axiomatizable;
(c) Each T_j knows that the simultaneous consistency of every agent implies the consistency of U, that is: $\vdash_{T_j} \wedge_{k=1,\ldots,m} Coer(T_k) \longrightarrow Coer(U)$, but not necessarily T_j can prove $Coer(U)$;
(d) U has among its models the standard model Z of $PRA(Z)$ (naturality condition);
(e) U includes transfinite induction only up to the ordinal $\beta^* \equiv \cap_{j=1,\ldots,m} \beta_j$, where β_j is the ordinal of transfinite induction possessed by the single T_j. Hence, U's proof-theoretic strength cannot be higher than that of any T_j.

Given our definition, we now take up the problem of existence of at least a system U of common rationality for S. The following theorem shows that, if in S there is a complex epistemic interaction of agreement, then (under rather mild conditions on the single T_j's - there is at least one system U of common rationality.

Theorem 1 *Let $S \equiv \{T_1, ..., T_m\}$ be a democratic society with agents at the level of inductive rationality. Assume that for all T_j in S the following holds:*
(i) T_j's inductive inference is given by sequent rules, and for all T_j's the rule in $R(T_j)$ are only those of $PRA(Z)$ and induction rules;
(ii) For every T_j, the proof-trees admit a normal form such that, given a model M of the axioms of Tj, in the conclusion of a rule instance in a proof P in T_j is M-true, then the premises are M-true;
(iii) T_j's axioms are such that in T_j there exist no proof of the consistency of systems, V say, which are extensions of $PRA(Z)$ not containing transfinite inductions up to V's ordinals.
If $\mathcal{B} \equiv \{B_1, ..., B_m\}$ is an m-ple of sentences such that $(T_1 + \{B_r\}_{r \neq 1}, ..., T_m + \{B_r\}_{r \neq m})$ is a complex epistemic interaction of agreement, then there exists a system of common rationality U, which can be constructed starting from the proofs of the statements generating the interaction.

Proof. Let $P_1, ..., P_m$ be respectively the proof-trees of theorems $B_1, ..., B_m$ in systems $T_1, ..., T_m$. By definition of complex epistemic interaction, every P_k, $k = 1, ..., m$, contains at least one axiom A_k of T_k which is not a theorem of any other T_j, $j \neq k$: indeed, if the root B_k of the proof P_k is complex, by definition the set $\mathcal{J}_k$ of induction rules in P_k has equal strength for all $k = 1, ..., m$. Hence, if T_k had as own theorems all the proper axioms required in the proof P_j, $j \neq k$, we would have $\vdash_{T_k} B_j$, contrary to the definition of epistemic interaction. Morevore, by the definition of epistemic interaction, there exists a common model M of the sentences $B_1, ..., B_m$, which is also an expansion of the standard model $\mathcal{Z}$ of $PRA(Z)$. By point (ii) above, there follows that all the axioms of all the trees $P_1, ..., P_m$ are true in M, and in particular so are all axioms A_k, $k = 1, ..., m$, selected by each P_k. Hence, all A_k are consistent with each other, and tre in the standard model $\mathcal{Z}$. Now define $A(U) \equiv \{A_1, ..., A_m\}$. Given homogeneity of a complex epistemic interaction, from every P_j one can obtain - within T_j - a proof of $Coer\,(PRA\,(Z)_k)$, for a common maximum finite k. Given our hypotheses, such a proof (and hence P_j) must contain a transfinite induction rule up to the ordinal of $PRA\,(Z)_k$, γ say. We accordingly define as the set of rules $R(U)$ of U, the set of induction rules $\{I_1, ..., I_m\}$ in which instances occur in each P_j, and which includes a transfinite induction up to γ. The system $PRA(Z) + Ord + A(U) + R(U)$ is consistent and preserves the standard model $\mathcal{Z}$, since the induction rules generate $\mathcal{Z}$-true conclusions from $\mathcal{Z}$-true premises. ■

We now want to discuss the relationship between the existence of a system of common rationality U for S, and the fact that some T_j's 'know' the rationality (consistency) or irrationality of other agents - technically, they can prove theorems like $Coer\,(T_k)$ or $\sim Coer\,(T_k)$ for some $k \neq j$

On the one hand, U is clearly incompatible with $\vdash_{T_j} Coer\,(T_k)$ for any $j \neq k$: U has been defined only for agents at the level of inductive rationality, the complexity (upper and lower) bounds of which prevent an agent from proving another's consistency. This is a substantial modeling choice, not a merely technical device. Indeed, a non trivial notion of common rationality requires that agents be all endowed with transfinite induction beyond ω; however, allowing for an agent being able to prove another's consistency would break the limit we have *a priori* established on the agents' cognitive power - e.g., by introducing induction beyond ε_0, or anyway by having, in the axiom sets, formulas which are true in the standard model and imply the consistency of PA.

On the other hand, admitting $\vdash_{T_j} \sim Coer\,(T_k)$ for some $j \neq k$ has no strong implications on the proof-theoretic strength of T_j. Rather, it would describe a non-standard relationship between T_j and 'truth' - i.e., a consistent, but highly subjective knowledge on T_j's part. The conflict between a commonly rational system U, and $\vdash_{T_j} \sim Coer\,(T_k)$ holding for some T_j, may be formally put in the following way: it can be shown that if indeed $\vdash_{T_j} \sim Coer\,(T_k)$, then T_j negates the existence of a system of common rationality, even though the latter exists. This will be shown Theorem 2, to introduce which, however, some preliminary work is necessary.

We can formalize in the common language of society S a statement of existence of a system of common rationality, as follows. Recall that, the language of $PA + Ord$ being suitable endowed with variables of second-order formulas, we are able to express the axiom and rule schemata of T_j in the formal language of T_j. Since T_j is recursively axiomatized, we can define the following recursive predicates

$$A_k\,(x) \leftrightarrow \text{“}x \text{ is the Gödel number of either an axiom of } T_k\text{, or an axiom schema of } T_k\text{”}$$

$$R_k\,(x) \leftrightarrow \text{“}x \text{ is the Gödel number of either a rule instance of } T_k\text{, or rule schema of } T_k\text{”}$$

The classical recursive predicate $Prov_{T_k}\,(m,n) \leftrightarrow$ “m is the Gödel number of a proof in T_k of the formula B, whose Gödel number is n” can then

be defined taking into account these extended notions of T_k-axioms and T_k-rules: the meaning of the standard provability sentence $\mathrm{Pr}_{T_k}(B)$ will be accordingly extended. We can also extend it to the case where the second argument is the Gödel number of a rule in the following way: $Prov_{T_k}(m,n) \leftrightarrow$"$m$ is the Gödel number of a T_k-proof Q, where (i) the premises of the rule $\mathfrak{R}$, whose Gödel number is n, occur as roots of Q-sub-proofs, (ii) the conclusion of $\mathfrak{R}$ occurs as a Q-root, and (iii) every branch of Q contains a premise of $\mathfrak{R}$": this makes it meaningful the sentence $\mathrm{Pr}_{T_k}(\mathfrak{R}) \leftrightarrow$"the rule $\mathfrak{R}$ is derivable in T_k".

We can now define the recursive function $Sist$: $Sist(x_1,...,x_r;y_1,...,y_s) \equiv$ "Gödel number of the deductive apparatus given by the axioms whose Gödel numbers are x_j, $j = 1,...,r$, and the inference rules whose Gödel numbers are y_j, $j = 1,...,s$". Also, one can define the recursive predicate $Prov^{\star}(x,y,z) \leftrightarrow$"$x$ is the Gödel number of a proof of the sentence whose Gödel number is z through the deductive apparatus whose Gödel number is y". Accordingly, the formula $\exists u Prov^{\star}(u, \ulcorner \perp \urcorner, Sist(x_1,...,x_r;y_1,...,y_s))$ has the meaning "The deductive apparatus whose Gödel number is $Sist(x_1,...,x_r;y_1,...,y_s)$ is consistent", which we shorten as $Coer(Sist(x_1,...,x_r;y_1,...,y_s))$. Moreover, the consistency between any recursively axiomatized system V and a deductive apparatus whose code is $Sist(x_1,...,x_r;y_1,...,y_s)$, can be so expressed: $\forall z[\mathrm{Pr}_V(z) \longrightarrow Coer(Sist(x_1,...,x_r;y_1,...,y_s))]$.

Given these premises, we can now define an *existence statement* for common rationality, as follows:

Definition 10 *A* statement of existence of a system of common rationality *for society S is the following sentence $Com(S)$:*
$Com(S) \equiv \exists x_1,...,\exists x_m \exists y_1,...,\exists y_m$
$[A_1(x_1) \wedge ... \wedge A_m(x_m) \wedge R_1(y_1) \wedge ... \wedge R_m(y_m) \wedge \wedge_{i=1,...,m} \sim \mathrm{Pr}_{PRA(Z)}(x_i) \wedge$
$\wedge_{i=1,...,m} \sim \mathrm{Pr}_{PRA(Z)}(y_i) \wedge \wedge_{i \neq 1} \sim \mathrm{Pr}_{T_i}(x_1) \wedge ... \wedge \wedge_{i \neq m} \sim \mathrm{Pr}_{T_i}(x_m) \wedge$
$\wedge_{i=1,...,m} \forall z [\mathrm{Pr}_{T_i}(z) \longrightarrow Coer(Sist(x_1,...,x_m;y_1,...,y_m))]]$

The following is noteworthy:

Remark 2 *if a system of common rationality $U(S)$ exists, then $Com(S)$ must be true in an expansion of the standard model.*

We can now prove the following:

Theorem 2 *(A) If, for any arbitrary* $j \in \{1, ..., m\}$, $\vdash_{T_j} \sim Coer\,(T_k)\,, k \neq j$, *then* $\vdash_{T_j} \sim Com(S)$, *i.e.* T_j *negates the existence of a system of common rationality.*
(B) The fact that $\vdash_{T_j} \sim Com(S)$, *is* not *sufficient to rule out that a system of common rationality* U *exist for* S *and that* $Com(S)$ *be true in an expansion of the standard model* $\mathcal{Z}$ *of* PA.

Proof. *(A)* From the basic properties of Predicate Calculus LK, we have that
$\vdash_{LK} \sim Com(S) \leftrightarrow \forall x_1, ..., \forall x_m \forall y_1, ..., \forall y_m [\sim A_1\,(x_1) \vee ... \vee \sim A_m\,(x_m) \vee \sim R_1\,(y_1) \vee .. \vee \sim R_m\,(y_m) \vee \vee_{i=1,...,m} \mathrm{Pr}_{PRA(Z)}\,(x_i) \vee \vee_{i=1,...,m} \mathrm{Pr}_{PRA(Z)}\,(y_i) \vee \vee_{i \neq 1} \mathrm{Pr}_{T_i}\,(x_1) \vee ... \vee \vee_{i \neq m} \mathrm{Pr}_{T_i}\,(x_m) \vee \vee_{i=1,...,m} \exists z [\mathrm{Pr}_{T_i}\,(z) \wedge Coer\,(Sist(x_1, ..., x_m; y_1, ..., y_m))]]$, shortened as $\forall x_1, ..., \forall x_m \forall y_1, ..., \forall y_m D(x_1, ..., x_m; y_1, ..., y_m)$.
Note $Prov^*(t, \ulcorner \perp \urcorner, Sist(x_1, ..., x_r, \ulcorner \perp \urcorner; y_1, ..., y_s))$ is by definition $PRA(Z)$-equivalent to the true $\perp$, whatever x_r, y_r and t. Since by assumption $\mathrm{Pr}_{T_k}\,(\ulcorner \perp \urcorner)$ is a T_j-theorem, and surely $\exists u Prov^*(u, \ulcorner \perp \urcorner, Sist(x_1, ..., x_r, \ulcorner \perp \urcorner; y_1, ..., y_s)) \equiv \sim Coer\,(Sist\,(x_1, ..., x_r, \ulcorner \perp \urcorner; y_1, . .., y_s))$, we get $\vdash_{T_j} \mathrm{Pr}_{T_k}\,(\ulcorner \perp \urcorner) \wedge \sim Coer\,(Sist\,(x_1, ..., x_r, \ulcorner \perp \urcorner; y_1, ..., y_s))$, whence, by the $\exists - \mathcal{R}$ rule, we get: $\vdash_{T_j} \exists z [\mathrm{Pr}_{T_k}\,(z) \wedge \sim Coer(Sist(x_1, .. ., x_r, \ulcorner \perp \urcorner; y_1, ..., y_s))]$, with x_r, y_r arbitrary terms which may be considered free variables. By the LK-rule $\vee - \mathcal{R}$, we have $\vdash_{T_j} \vee_{i=1,...,m} \exists z [\mathrm{Pr}_{T_k}\,(z) \wedge \sim Coer(Sist(x_1, ..., x_r, \ulcorner \perp \urcorner; y_1, ..., y_s))]$, and, similarly, $\vdash_{T_j} D(x_1, ..., x_m; y_1, ..., y_m)$, with x_i, y_i free variables. Hence, with a $\forall - \mathcal{R}$ rule we have $\vdash_{T_j} \forall x_1, ..., \forall x_m \; \forall y_1, ..., \forall y_m \; D(x_1, ..., x_m; y_1, ..., y_m)$, and hence $\vdash_{T_j} \sim Com\,(S)$.
(B) Since all T_i's are consistent, T_j lacks the standard model for his theorem $\mathrm{Pr}_{T_k}\,(\ulcorner \perp \urcorner)$: hence it is endowed with non-standard models only, and $\sim Com\,(S)$ is true in all these models.However, $Com(S)$ being complex, it is simultaneously possible that it be true in an expansion of t he standard model $\mathcal{Z}$, and that a system of common rationality exist.Consider the following example:
$S = \{T_1, T_2\}$
$T_1 = PRA(Z) + Ord + I(\alpha) + Coer\,(PRA\,(Z)_{18}) + RI(\gamma)$
$T_2 = PRA(Z) + Ord + I(\beta) + Coer\,(PRA\,(Z)_{15}) + RI(\lambda) + \mathrm{Pr}_{T_1}\,(\perp)$,
with $I(\alpha), I(\beta)$ axioms of transfinite induction up to α, β; $RI(\gamma)$ and $RI(\lambda)$ rules of transfinite induction up toγ,λ; α, β, γ, λ are ordinals below ε_0 and greater than ω, and such that none of the corresponding inductions is sufficient to prove $Coer\,(PRA\,(Z)_{18})$ and $Coer\,(PRA\,(Z)_{15})$. Let α be the minimum of α, β, γ, λ. T_2 is consistent, since $PRA(Z)+I(\beta)+Coer\,(PRA\,(Z)_{15})$ does not prove the consistency of T_1; hence a system U of common rational-

ity for S has proper axioms $A(U) = \{I(\alpha), Coer\,(PRA\,(Z)_{15})\}$, and rules $R(U) = \{RI(\alpha)\}$. ■

5 Concluding remarks

In this paper we have proposed a formalization of knowledge, common knowledge and common rationality based on treating agents as consistent first-order formal systems. A crucial point of our approach has been the formalization of knowledge, starting from the modal system G, which is canonically isomorphic to the standard Provability Logic of Arithmetic. A methodological contribution of the paper is the introduction of proof theory as a basis for studying epistemic interaction.

The next point on our resarech agenda is exploring the relationship between a system of common rationality lke U, and that part of individual rationality which does not contribute to common rationality. These should deliver a notion of equilibrium between individual and common rationality, which we see as preliminary to many equilibrium notions currently in use.

References

[1] Bairwise, J. (1977): An Introduction to First-Order Logic, in J.Bairwise (ed), Handbook of Mathematical Logic, North Holland, Amsterdam (6th printing: 1991), 5-46.

[2] Boolos, G. (1979): The Unprovability of Consistency, Cambridge University Press, Cambridge.

[3] Chang, C. and J.Keisler (1973): Modal Theory, North Holland, Amsterdam.

[4] Geanakoplos, J. (1994): Common Knowledge, in R.J.Aumann and S.Hart (eds), Handbook of Game Theory with Economic Applications, Elsevier, Amsterdam, 1438-95.

[5] Gentilini, P. (1992): Provability Logic in the Gentzen Formulation of Arithmetics, Z.Math.Log.Grund.Math., 38, 536-50.

[6] Gentilini, P. (1998): Proof-Theoretic Modal PA-completeness, Studia Logica, forthcoming.