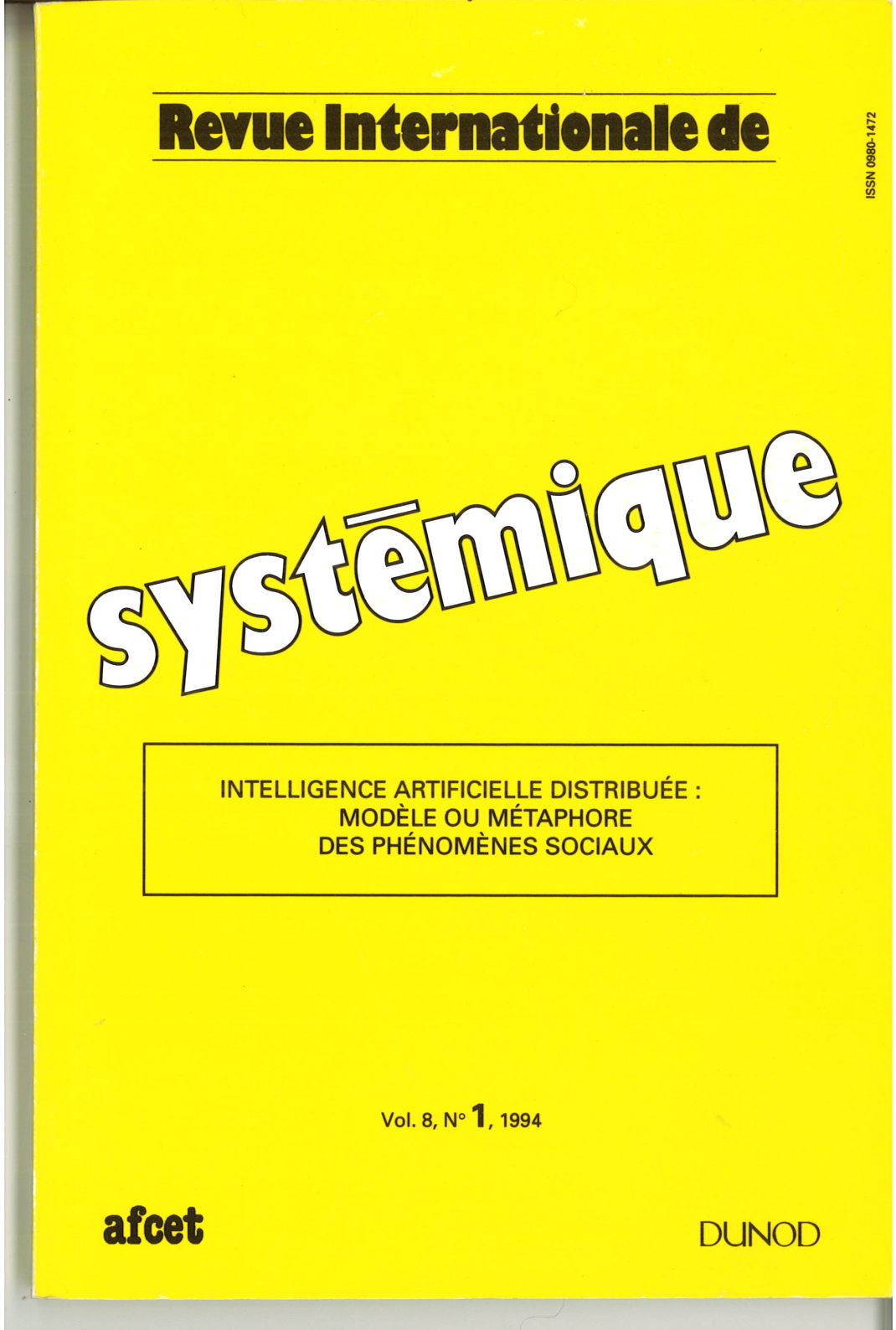

AFSCET

Revue Internationale de systēmique

Revue
Internationale
de Sytémique

volume 08, numéro 1, pages 123 - 132, 1994

The logical Way if Describing Societies

Jacques-Paul Dubucs

Numérisation Afscet, janvier 2016.

REVUE INTERNATIONALE DE SYSTÉMIQUE
Vol. 8, N° 1, 1994, pp. 123 à 132

THE LOGICAL WAY OF DESCRIBING SOCIETIES

Jacques-Paul DUBUCS [1]

Résumé
Une théorie correcte de la vie sociale doit pouvoir expliquer en quoi consiste la capacité manifestée par chaque individu de former une représentation interne de lui-même et des autres. Compte-tenu de phénomènes d'« opacité référentielle » bien connus, de tels systèmes représentationnels ne peuvent être décrits dans le format de la logique extensionnelle classique. Ils peuvent l'être, en revanche, dans le cadre d'une logique « intensionnelle » dans laquelle la référence d'une expression dépend du « sens » – et pas seulement de la référence – de ces constituants. L'article montre comment l'on peut définir de cette manière la hiérarchie des connaissances dans un système multi-agents.

Abstract
A correct theory of social life has to explain the general ability to build and use an internal representation on oneself and the others. Viewing well-known "referential opacity" phenomena such representational systems cannot be described in the frame of classical (extensional) logic. But they can be analysed in an "intensional" setting, in which the reference of an expression depends on the "meaning" – not only on the reference – of its component parts. The paper shows how one can in this way define the hierarchy of knowledge in a multi-agents system.

Let us begin with a commonplace. Social life rests on the general ability to build and use an internal representation of oneself and the others. Thus the representation each agent builds of his environment has to include a representation of the representation his partners themselves build of their own environment. We shall give here some examples of this kind of *embedded representations*.

1) According to Premack's work in ethology, a female chimpanzee who wants to divert the attention of her regular companion will utter an alarm cry, in order to modify the representation he has of his environment. She has therefore some representation of his own representation.

1. Institut d'Histoire et de Philosophie des Sciences et des Techniques, 13, rue du Four, 75006 Paris.

Rev. intern. systémique. 0980-1472 Vol. 8/94/01/ $ 4.00/

2) Look to the cognitive machinery that underlies driving behavior. Why do I not give way to the vehicles coming from the left? Not only to be sure, because I know that the law allows me to do that: I do not feel inclined to be injured, even within my rights. I behave this way because I have good reasons to think that the other drivers know that the cars coming from the right have priority.

3) Economic behavior is another major source of examples of embedded knowledge or belief. If I am rational, my Stock Exchange transactions should not be grounded on my opinions about the "real" value of the stocks, but on my opinions about the opinions of the others about this value, for that is their opinions which actually determine the price of the stocks. But the others are in the same situation as me. If they are rational, their opinions are also second-order opinions about the opinions of the others, myself included, in such a way that my actual opinions are third-order opinions, a.s.o.

4) According to Grice (1968), a speaker who utters the sentence S to mean that p is prompted by a complex of intentions. To be sure, one of them, say I1, is the intention of making his audience believing that p. But he also has the intention I2 that the very reason for the hearer of believing that p should be the recognition by him of the intention I1...

All these examples make clear the necessity, for the theorist interested in the study of social life, to have a correct theory of embedded (or higher-order) representations. Logic can provide the general framework in which such a theory may be expressed. The aim of the present paper is to expose the rudiments of such a framework.

I. TECHNICAL BACKGROUND: LOGIC FOR DESCRIBING REPRESENTATIONAL SYSTEMS

The leading principle of modern logic since Frege is the *compositionality principle* (C.P.). According to C.P., the analysis of any language is commited to satisfy the two following requirements:

a) The complex expressions of L have to be considered as made of elementary constituents by successive applications of a finite number of morphological rules.

b) The semantical value of a complex expression of L has to depend only on the semantical value of its constituents, and it has to be calculable on this basis by successive applications of a finite number of valuation rules that work like the morphological rules do.

To sum up: whenever some morphological rule allows to build some well-formed expression $*(x_1,\ldots,x_n)$ from admittedly well-formed expressions $x_1, \ldots, x_n$, there must be some semantical rule f_* to determine the value of the complex expression on the basis of the value of its constituents:

(C.P.) $$\mathrm{Val}\,(*(x_1,\ldots,x_n)) = f_*(\mathrm{Val}(x_1),\ldots,\mathrm{Val}\,(x_n))$$

(C.P. is the only principle able to explain how a system equipped with finite cognitive ressources can however succeed in mastering languages that potentially countain infinitely many sentences.)

I.1. Extensional logic

Logicians have been first engaged in the study of the language of mathematics, where C.P. takes a very simple form. Here the semantical value of a designator ("4", "the maximum of the function f on the interval I") is simply what the designator designates; and the value of a complete sentence is simply its truth-value. Thus C.P. takes the form *extensionality principle*:

(E.P.) $$\mathrm{Ref}\,(*(x_1,\ldots,x_n)) = f_*(\mathrm{Ref}\,(x_1),\ldots,\mathrm{Ref}\,(x_n))$$

This principle states that the truth-value of a complex sentence depends *only* on the truth-value of its subsentences, and that it is therefore not affected by the "meaning" of them. E.g. we will have

$$\mathrm{Ref}\,(\mathrm{A}\ \&\ \mathrm{B}) = f_{\&}(\mathrm{Ref}\,(\mathrm{A}),\ \mathrm{Ref}\,(\mathrm{B})), \qquad (1)$$

where $f_{\&}$ is defined by the well-known truth-table $f_{\&}(a,b) = \mathrm{TRUE}$ if $a = \mathrm{TRUE}$ and $b = \mathrm{TRUE}$, and $f_{\&}(a,b) = \mathrm{FALSE}$ if not.

E.P. states also that the truth-value of a sentence is never affected by the repleacement of a designator by a co-referential designator:

$$\text{If } \mathrm{Ref}(u) = \mathrm{Ref}(v), \text{ then } \mathrm{Ref}\,(\mathrm{A}[\ldots u \ldots]) = \mathrm{Ref}\,(\mathrm{A}[\ldots v \ldots]). \qquad (2)$$

We have thus to do with a very coarse taxonomy, that distinguishes neither between designations of the same object, nor between sentences with the same truth-value. A so rudimentary classification is however sufficient for the language of "ordinary science": e.g. this extensional frame has been very successfully applied in the early 20th century to the problems of the foundations of mathematics. But it is by no way fine enough if we intend to describe representational systems.

I.2. Intentional logic

An intelligent agent acts on the basis of representations. I do not mean that intelligent agents make an exception to the natural laws, nor that the behavior of them does not have any physical causes. But that the regularities of this behavior cannot be grasped at the level of physical causality: this regularities only appear by reference to *representational states* of the agent, who may be disposed to act in a certain way towards some object of his environment only if he apprehends this object under such and such description (but not under such other description).

In these conditions E.P. becomes inapplicable, as clearly showed by the following counter-example of *referential opacity*: despite the inference

(*a*)
$$\frac{\text{Œdipus intended to kill Laïos}\qquad \text{Laïos = Œdipus father}}{\text{Œdipus intended to kill father}}$$

is recommanded by (E.P.), it has to be considered as incorrect (at least if one wants to understand anything at all in Sophocles tragedy...).

So far as the analysis of representational systems is concerned, E.P. has therefore to be removed. In order however to keep C.P. alive, we have of course to modify the current notion of semantical value: it has to be defined as a function of the "meaning" – not only of the "reference" of the expressions. The basic idea of the intensional logic is to determine the meaning by a simple *relativization* of reference: the meaning (or the "intension") of an expression is given by its reference in every possible world w (because to understand a sentence, to grasp its meaning, is to be able to say in what circumstances it would be true or false). C.P. takes now the form of the *intensionality principle*:

(I.P.) $$\mathrm{Ref}_w(^*\langle x_1, \ldots, x_n\rangle) = f^*[w, \mathrm{Ref}_{w\in W}(x_1), \ldots, \mathrm{Ref}_{w\in W}(x_n)].$$

Now the truth-value of a sentence in the "real" world may clearly depend on the truth-value of its subsentences in other possible worlds. In our example I.P. applies as follows: the truth-value of "Œdipus wants [Laïos is killed]" may differ from the truth-value of 'Œdipus wants [Œdipus father is killed]", though the embedded subsentences are equivalent in the real world: one only needs that these subsentences take distinct values in some other possible worlds (possible worlds in which Laïos just differs from Œdipus father). To sum up, we shall represent mental states as "to want to p", "to know that p", a.s.o. by the class of the possible worlds where p is satisfied.

More technically, we may give the following semantical analysis of such a sentence as "a knows that A" (in symbols: "KaA"). Given a "Kripke structure" where W is a non-empty set of "possible worlds", an "accessibility relation" between these worlds, and a valuation, we define the truth-value of KaA in the world w by:

(3)

For example the following diagram (*figure* 1) represents the case where the agent a know that p but ignores that q:

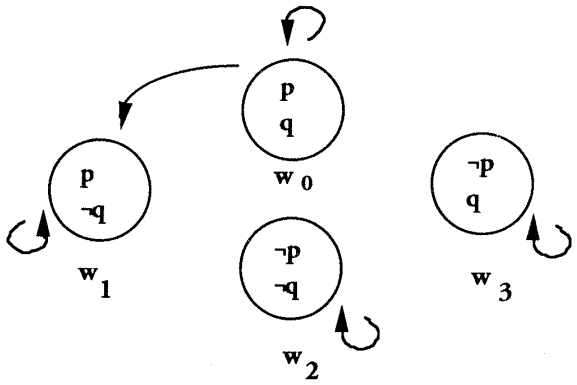

Figure 1.

Note that the very content of the knowledge of an agent is expressed by the accessibility relation: from an intuitive viewpoint, w' is accessible from w means "so far as a (in w) knows, w' could be the right description of w". Therefore the accessibility relation which characterizes a fully ignorant (resp. omniscient) agent is W X W (resp. Diag (W) =).

The set of the formulas which are valid in any Kripke structure is axiomatised by the basic system K, which conctains all the propositional tautologies and all the formulas of the kind

(K) $$\mathrm{K}a(\mathrm{A} \rightarrow \mathrm{B}) \rightarrow (\mathrm{K}a\mathrm{A} \rightarrow \mathrm{K}a\mathrm{B})$$

and which admits the rules of modus ponens A, A → B/B and necessitation. Three supplementary axioms are arguably required in order to capture the full

meaning of the notion of knowledge, namely:

(T)	$KaA \rightarrow A$	(what is known is true)
(S_4)	$KaA \rightarrow KaKaA$	(positive introspection)
(S_5)	$\neg KaA \rightarrow Ka\neg KaA$	(negative introspection)

(to these axioms correspond three algebraic conditions on the relation R, respectively: reflexivity, transitivity and euclidianity).

It is worth to notice that *belief* may be similarly characterized in terms of possible worlds (the operator B*a* ("*a* believes that") is arguably definable from the previous system by removing the axiom T). More generally, any *propositional attitude, i.e.* any mental state expressible by a sentence as "the agent *a* ... that *p*', may be represented in this frame.

II. EXPLAINING ACTION IN THE POSSIBLE-WORLDS FORMAT

Action is not explainable by reliefs alone, but by beliefs and desires together (or, in terms more familiar to the economists, by probabilities and utilities together). Both components are linked by the *pragmatic principle:* agents undertake just the actions that, according to their beliefs, will lead to the satisfaction of their desires. This principle is appositely expressible in the possible-wolds format.

One's beliefs delineate a part of the set W of possible worlds, namely the part B containing the worlds compatible with the content of these beliefs. Similarly, one's desires delineate the subset D of the worlds in which they are satisfied. Now an action may be viewed as a transformation A of W: A(*w*) is intended to be the world that results from doing the action A in the world *w*. An action A may be termed *optimal* for an agent X if and only if it transforms each world compatible with X's beliefs into a world compatible with X's desires, i.e. if and only if

$$(O) \qquad \forall w \in W\,[w \in B \rightarrow A(w) \in D]$$

The pragmatic principle affirms that human action is always optimal in this sense. The common sense *explanation* of behavior along this principle is thus at the same time a *normalization* of it: according to this view, the cognitive states (beliefs, desires) that are the reasons of an action are just the states with respect to which the action appears as optimal.

The current theory of economical behavior works basically along the same line. It supposes that the agents are guided by the principle of maximization of expected utility, which asserts that between several possible actions, one has to choose the action A for which the term

$$\sum_{w \in W} pr(w) \cdot u(A(w))$$

takes the greatest value, where $pr(w)$ and $u(w)$ are the subjective probability and utility attached to the eventuality *w*. But this assumption is a simple generalization of the pragmatic principle to the case of partial beliefs and graded desires. For if beliefs and desires are perfectly categoric, *i.e.* if the range of values of *pr* and *u* is the pair {0, 1} instead of the whole interval [0, 1], then these functions may be considered as the respective characteristic functions of B and D, in such a way that the expected utility of A reduces to

$$\sum_{w \in B} u(A(w)),$$

term which attains its maximum when (O) is satisfied.

III. MULTI-EPISTEMIC LOGIC

We are now ready to expose the basic notions of a logic intended to formalize the epistemic interactions in a *society of cognitive agents.* This so-called multi-epistemic logic arises from the epistemic logic above by mere generalization to n agents. In the formula K*a* A, the symbol K*a* is no longer indivisible, but it has the status of an indexed modality. The relevant semantical structure is , where W and V are as above, and R_i the accessibility relation that characterizes the *i*-th agent. Several interesting notions are definable in this framework (Table 1):

Table 1. Hierarchy of knowledge in a multi-agent system G

	Omniscience	Diag (W)
I_GA	Implicit knowledge	$\bigcap_{i \in G} R_i$
KiA	Individual knowledge	R_i
U_GA	Universal knowledge	$\bigcup_{i \in G} R_i$
C_GA	Common knowledge	$Cl(\bigcup_{i \in G} R_i)$
	Full ignorance	WXW

Here some comments.

1) *Implicit knowledge* is the knowledge the members of the society would have if they cooperated. For example if *a*, who knows that *p*, and *b*, who knows that *p* implies *q*, exchanged their information, then clearly both agents would obtain the information *q*. From a formal point of view: if there is a world that is known by some member *i* of G to be impossible (in such a way that this world cannot be reached via R_i), then this world cannot be considered as possible by any member of G after the cooperation process has drawn to end.

2) *Universal knowledge* is the knowledge all the members of the society possess. This notion, which is clearly dual to the previous one, does not deserve many comments.

3) *Common knowledge* is by far the most interesting concept, both for the width of its applications in social sciences and by the technical problems it raises, and it deserves more substantial comments.

As a cement of social life, universal knowledge is not enough. As we have said in the above introduction, many examples show the necessity of continuing the analyses further by introducing *embedded knowledge* (knowledge that somebody knows that p, knowledge that everybody knows that everybody knows that, a.s.o.).

We obtain on this way second-order universal knowledge U_G^2, defined by $U_G^2 A = U_G U_G A$, and more generally *n* th-order universal knowledge $U_G^n = U_G U_G^{n-1}$. Note that the last notion has a very natural semantical counterpart: if we define $S = \bigcup_{i \in G} R_i$, this counterpart is the relation S^n on W defined by:

$$S^1 = S$$

and

$$w S^n w' \text{ iff } \exists w''\left(w S^{n-1} w'' S w'\right).$$

In the following example, we have (in w_0) second-order universal knowledge of *p*, but not of *q*, for *b* does not know that a knows that *q* (*b* envisages as a real possibility the world w_1, in which a does not know that q). Notice that w_2 is related by S^2 to W_0, *i.e.* that we can move from w_0 to w_2 in stages by following either *a*'s or *b*'s arrows (figure 2).

One can however argue that universal knowledge of *finite* order is not enough either. For if we take seriously the mirror situation we have described in the third example of the introduction, we are fatally driven to consider *infinitely embedded*

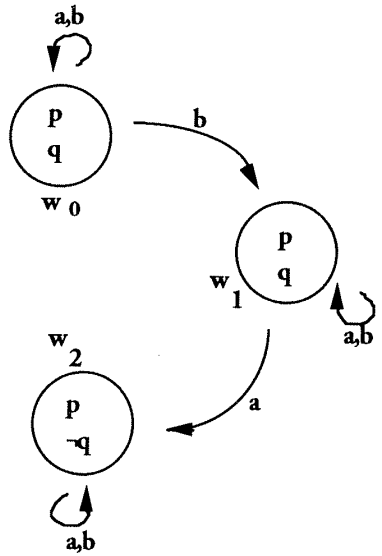

Figure 2

knowledge or belief. How to cope with this infinity? From a semantical point of view, it suffices to consider the relation on W defined by the transitive closure $Cl\left(\bigcup_{i \in G} R_i\right) = \bigcup_{n>0} S^n$. But there are competing views on the very definition of common knowledge. The most used notion is *iterative*: it is defined, in the case of two agents, by: *a* knows A, and *b* knows A, and *a* knows that *b* knows A, and *b* knows that *a* knows A, a.s.o.. Viewing the infinity of this conjunction, it is preferable to define common knowledge by means of some kind of *circularity* (fixed-point account), as argued by Barwise (1989).

IV. TOWARDS A HYPER-INTENSIONAL LOGIC

The intensional logic, which equates the semantical value of an expression with the class of its referents in all possible worlds (not only the actual world) provides a taxonomy of representations much finer than the classical, extensional logic. But it is committed to treat as indiscernable two expressions that have the same reference in any possible word (that means: provably equivalent), and particularly two sentences that are true in any possible world (that means: logically true). In

other words, we are committed, in the format of intensional logic, to suppose that cognitive agents are always able to consider as identical two provably equivalent representations: they are supposed to be "logically omniscient". But, to be sure, this hypothesis is crudely irrealistic, and the resulting modelisation is certainly incorrect. One needs therefore a finer logical analysis. There are today a lot of attempts to *locally* remedy this defect of intensional logic (e.g.: addition of "impossible possible worlds" beside standard worlds). But no one is very convincing, and we have probably to confess that, despite the precision it has provided in the description of the cognitive mechanisms underlying social life, possible-words semantics is today a dead end: the logical principles of a correct theory of cognitive representations are still to be found.

References

J. BARWISE, On the Model Theory of Common Knowledge, in *Situation Theory and Related Topics*, CSLI, Stanford, 1989.

J. DUBUCS, The Problem of Logical Omniscience, in *Logique et Analyse*, n° 133-134 *(International Symposium ou Epislemic Logic)*, 1991, pp. 41-55.

H. P. GRICE, Logic and Conversation, in P. COLE, J. MORGAN (Eds.), *Syntax and Semantics 3: Speech Acts*, Academic Press, 1975.

J. Y. HALPERN, Y. MOSES, *Knowledge and Common Knowledge in a Distributed Environment,* IBM Research Laboratory, San José, CA, RJ 4421, 1984.

L. LISMONT, *La connaissance commune. Approches modale, ensembliste et probabiliste*, Thèse de Mathématiques, Université de Louvain, 1992.

Ph. MONGIN (Ed.), Epistemic Logic, special issue of *Theory and Decision*, in press.