# Towards a First-Order Approach for Social Agents: Preliminary Report

**Maria Fasli**

University of Essex, Department of Computer Science
Wivenhoe Park, Colchester CO4 3SQ, UK
mfasli@essex.ac.uk

**Abstract**

The study of intelligent agents capable of rational as well as social behaviour has received an increasing attention within Computer Science and in Artificial Intelligence in particular. A theory of agents can be used as a specification language in order to design, build, study and verify multi-agents systems. In this paper after arguing about the importance of common knowledge in multi-agent systems, we provide a method for the formal description of such systems. We present a first-order self-referential framework for reasoning about truth and modal knowledge and common knowledge. We then look at an extension of the basic logical machinery in which knowledge and common knowledge are predicative modalities. We continue by discussing a special case, a well-known logical paradox, the surprise examination, which involves self-reference in a multi-agent domain and we employ common knowledge to investigate it.

## Introduction

The term autonomous agent is usually employed in Computer Science and Artificial Intelligence to describe computer systems or programs that are capable of independent action and rational behaviour in an open, and often unpredictable environment. A multi-agent system is a collection of such intelligent agents that continuously interact with the environment and each other. Therefore they often need, to communicate, co-ordinate and collaborate with one another in order to achieve a mutual goal, perform a complex task or share resources. Here we adopt a mentalistic view for the agents, we define agents as having certain mental qualities or attitudes such as knowledge and beliefs and thus we characterise their behaviour in terms of these mental attitudes. This approach of ascribing human properties and attitudes to artificial agents, known as the intentional stance (Dennet 1987), is a useful and convenient means of describing complex systems, explaining and predicting their behaviour. Using the intentional notions we can formulate agent theories in order to describe an agent, its properties and its reasoning. A theory of agents can be viewed as a specification language and can be used to design, build, study and verify multi-agent systems.

For agents to interact not only with their environment but with the other agents as well, their theory should involve reasoning about the other agents and their knowledge, desires, goals, etc.. The first requirement that emerges for such a theory of agents is that it should be able to express self-referential statements. This stems from the fact that an agent's assertions about the other agents' knowledge may be self-referential. Self-reference occurs when an agent knows or believes a proposition about the knowledge or beliefs of other agents which in turn make reference to this very proposition. For instance:
A. I know all the facts that agent B knows
B. I know a fact that agent A knows.
As it is obvious agent's A knowledge refers to agent's B knowledge and vice versa. Although these seem perfectly sensible assertions, there are others, like: the "Liar": "This sentence is false" that can lead to inconsistencies.
Since we are interested in agents that interact with each other the second issue that should be taken into account is that of common knowledge. The concept of common knowledge is based on what everyone in a group of agents knows. For instance everyone in our society knows (ideally) that a green light means go and a red light means stop (Fagin et al.1995). This fact is also common knowledge, because not only everyone knows it, but everyone knows that everyone else knows it, and everyone knows that everyone knows that every individual knows it, and so on. Common knowledge was first studied in the context of conventions (Lewis 1969) but seems to be essential in co-ordination, collaboration, reaching agreements and in discourse understanding as well. Some of these issues will be discussed in the sequel.
In this paper we present an alternative possible worlds formalism for reasoning agents. We provide a first-order self-referential approach for reasoning about knowledge and truth in a multi-agent domain working within the framework for syntactic modalities of (Turner 1990). However, we allow multiple syntactic modalities for knowledge since we want the theory to be applied in a multi-agent domain and furthermore we incorporate new ones in order to capture the notion of common knowledge and its properties. Our main aim is to present an expressive and consistent self-referential framework for conceptualising the notions of truth, knowledge and

common knowledge. However, due to space limitations we do not present the model theory and the theorems are stated without proof. The structure of the paper which consists of four subsequent sections is as follows. In the next section we informally establish the necessity of common knowledge in a multi-agent system. The following section firstly describes the basic logical machinery that supports the notions of truth and modal knowledge and common knowledge and then an extension of the framework in which knowledge and common knowledge are treated as syntactic modalities. We continue by discussing a special case, the surprise examination, using common knowledge as the new means of investigating it. A summary and a brief discussion of our findings is then provided and the paper ends with the conclusions and a pointer to future work.

## The Need for Common Knowledge

In this section we will embark on a more detailed discussion of the concept of common knowledge and we will argue about its need in a multi-agent system. Recall from the discussion above that common knowledge is based on what everyone in a group of agents knows and that it was first discussed by (Lewis 1969) in his philosophical study of conventions. Although we are not going to engage in a sociological analysis of conventions and how they are formed, nevertheless we have to mention here that conventions and social norms are regularities in behaviour. A convention is nothing more than a group prescription to do or not to do given actions under certain circumstances, it is a co-operative social behaviour and everyone belonging to a certain group or community has to conform to it. What is important in the formation of conventions and social norms however, is the concept of common knowledge. In order for something to be a convention in a group, everyone must be aware of the conditions of the convention, and everyone must know that every other member of the group knows that and that everyone is willing to conform to this convention. In other words everyone in the group must have common knowledge of the convention.

Structured groups of human agents like societies, communities and organisations, operate according to certain predetermined authority relationships and social norms. Once an agent enters into a community he is given a specific role which entails a set of rights and obligations. Each member of the group knows its place and acts accordingly and furthermore each knows the implications of exercising rights and braking commitments. These authority mechanisms rely on the obedience of the addressees and their behaviour, and enable the whole group to act effectively and efficiently. In such a structured community the notion of common knowledge naturally arises. For instance, the leader of such a group has a certain predetermined authority and this is common knowledge among all the members of the group. All members also know that they are obliged to carry out certain tasks and obey the leader and this is profoundly common knowledge. However it seems reasonable to suggest that agents in artificial societies or in multi-agent systems with build-in or with the ability to form, authority relationships and groups should have common knowledge of the whole structure, as well of their role in it, and the rights and obligations regarding themselves and the other agents.

An ordinary way of acquiring common knowledge is through communication. Announcements made in public are sources of common knowledge. The purpose of the speaker is to make her intentions, goals, preferences etc. known to a group of other agents as a whole, that is she provides common knowledge. The content of the announcement can then be used by each of the members individually and as a group as well, in order to make decisions, plans or take action. Simple speech acts between two or more agents, exchanging information can be the source of common knowledge as well, on the grounds that the agent supplying common knowledge is trustworthy and reliable. From now on and throughout this paper we assume that we are dealing with rational, truthful and sincere agents, who have no intention of deceiving one another or supply false information. Common knowledge is also relevant in discourse understanding and in the use of a specific language among a population since grammar rules and word meaning are considered common knowledge. Consider the case of a new expression added to the language. In order to use this expression a group must have common knowledge of it, how it is used, what it means and that everyone knows its meaning and how to use it as well.

In a number of applications such as agreements, co-ordination and collaboration, common knowledge seems to be relevant as well. For instance suppose that two agents A and B need to agree on some statement $x$. It is reasonable to suggest that if the two agents agree on $x$ then each of them knows that they have agreed on $x$. This is a key property of agreement: each of the agents must know about the agreement and he must also know that every other participant in the agreement knows about it. Thus agreements always presuppose common knowledge among all the agents involved. In co-ordination situations each agent needs common knowledge for example of the schedule of actions to be performed. In cases such as the co-ordinated attack where common knowledge cannot be attained, co-ordination becomes difficult if not at all impossible (Fagin et al.1995).

But how does common knowledge affect an agent's social behaviour and influences an agent's decisions, future actions and goals? For instance, in the case where an agent belongs to a certain structured group, community or organisation, the agent has to adopt social commitments and must undertake obligations and rights. But a socially committed agent to a group or another agent, loses some of its autonomy in order to adapt to the organisational constraints. Therefore an agent having now common knowledge of certain restrictions may need to modify his

behaviour to conform to these restrictions. His goals and future plans and actions may be in conflict with the constraints and requirements of the group and thus may need modifications and alterations. Perhaps some of them need to be abandoned completely. Common knowledge is not an abstract definition but a key concept, a basic and active ingredient of an agent's reasoning process and affects and guides the agent's social behaviour .

## The Logical Framework

In this section we present an alternative possible worlds formalism for agents. Our logical machinery is based on the framework proposed in (Turner 1990) which we extend further by adopting the approach of (Fagin et al.1995) towards common knowledge.

Our language $\mathcal{L}$ is based on a First Order language and apart from the standard connectives and quantifiers it also includes: a) Three predicate symbols T, F and = for truth, falsity and equality respectively. b) Three modal operators $K_i$, $E_x$, $C_x$, for what an agent knows, what everyone knows and what is common knowledge in a group of agents. c) A set of variables V and a set of predicate symbols P.

Terms are: i) Constants, ii) Variables, iii) wffs can be treated as terms in the language in order to allow circular reference. Thus if $A(y_1, y_2,..,y_n)$ is a wff and $y_1, y_2,..,y_n$ are free variables then $(tA,y_1, y_2,..,y_n)$ is a term. (In what follows we write $A$ where no confusion can arise.)

Inductive definition of wffs:

i)If t and t' are terms then t= t', T(t) and F(t) are wffs

ii)If $x$ is a variable and A and B are wffs, so are $\neg A$, $A\wedge B$, $A\vee B$, $A\Leftrightarrow B$, $A\Rightarrow B$, $\forall xA$, and $\exists xA$

iii)If $A$ is a wff then $K_i(A)$, $E_X(A)$ and $C_X(A)$ are wffs.

A model for the logical language $\mathcal{L}$ is a tuple $\mathcal{M}=<\mathcal{W},K_i,\mathcal{D},T,F>$ where: $\mathcal{W}$ is a set of possible worlds, $K_i$, is a set of accessibility relations, $\mathcal{D}$ is the multi-agent domain of individuals, T is the extension of the truth predicate (that is T: $\mathcal{D}\times\mathcal{W}\rightarrow\{0,1\}$) and F is similarly the extension of the false predicate. We require the domain of individuals to be constant and Cartesian closed, and each individual constant to be a rigid designator.

The weakest logic that we can have for Truth, modal knowledge and common knowledge is $D_{Tm}$ :

k. $T(A\Rightarrow B)\Rightarrow(T(A)\Rightarrow T(B))$

d. $T(A)\Rightarrow\neg T(\neg A)$

bar $\forall xT(A)\Rightarrow T(\forall xA)$

Nec if $D_{Tm}\vdash A$ then $D_{Tm}\vdash T(A)$

K. $K_i(A\Rightarrow B)\Rightarrow(K_i(A)\Rightarrow K_i(B))$

D. $K_i(A)\Rightarrow\neg K_i(\neg A)$

BAR $\forall xK_i(A)\Rightarrow K_i(\forall xA)$

NEC if $D_{Tm}\vdash A$ then $D_{Tm}\vdash K_i(A)$

E. $E_X(A)\Leftrightarrow\bigwedge_{i\in X}K_i(A)$

C. $C_X(A)\Leftrightarrow E_X^k(A)$ for $k$=1,2,…..

IR. If $A\Rightarrow E_X(A\wedge B)$ then $A\Rightarrow C_X(B)$

We can strengthen the logic in two ways i) by adding stronger axioms for Truth and therefore weaken the necessitation rule in order to obtain consistency, or ii) by adding stronger axioms for modal knowledge. Let X be any subset of the standard axioms for truth {t,s4}:

t. $T(A)\Rightarrow A$

s4. $T(A)\Rightarrow T(T(A))$

Let Y be any subset of the standard modal axioms for knowledge {T,S4,S5} given below:

T. $K_i(A)\Rightarrow A$

S4. $K_i(A)\Rightarrow K_i(K_i(A))$

S5. $\neg K_i(A)\Rightarrow K_i(\neg K_i(A))$

Let us call the logic D[X,Y] the logic which comprises of the aforementioned $D_{Tm}$ logic and the sets of axioms Y and X together with the rules:

Neca. if $D[Y]\vdash A$ then $D[X,Y]\vdash T(A)$

Necb. if $D[X,Y]\vdash A$ then $D[X,Y]\vdash K_i(A)$

**Theorem 1.**

*D[X,Y] logics are consistent systems of truth and modal knowledge and common knowledge.*

Let us extend the language $\mathcal{L}$ to $\mathcal{L}1$ by adding three new predicates $KNOW_i$, $EK_X$ and $CK_X$ for expressing knowledge, what everyone in a group knows and common knowledge respectively. Since it is well known that a direct treatment of knowledge as a predicate can result in inconsistency, we are going to use an alternative approach here in order to obtain consistency on the one hand, and strong enough logics on the other. In this approach, following (Turner 1990) the knowledge predicate is defined as follows: $KNOW_i(A) =_{def} K_i(T(A))$

Now the central question is what are the available logics to us. The weakest logic is $D_{TK}$:

k. $T(A\Rightarrow B)\Rightarrow(T(A)\Rightarrow T(B))$

d. $T(A)\Rightarrow\neg T(\neg A)$

bar $\forall xT(A)\Rightarrow T(\forall xA)$

Nec if $D_{TK}\vdash A$ then $D_{TK}\vdash T(A)$

K. $KNOW_i(A\Rightarrow B)\Rightarrow(KNOW_i(A)\Rightarrow KNOW_i(B))$

D. $KNOW_i(A)\Rightarrow\neg KNOW_i(\neg A)$

BAR $\forall xKNOW_i(A)\Rightarrow KNOW_i(\forall xA)$

NEC if $D_{TK}\vdash A$ then $D_{TK}\vdash KNOW_i(A)$

Let X be the same set as in the simple modal case and Y be any subset of the axioms for knowledge {T,S4}:

T. $KNOW_i(A)\Rightarrow A$

S4. $KNOW_i(A)\Rightarrow KNOW_i(KNOW_i(A))$

Let us call $D_{TK}[X,Y]$ the logic which comprises of the aforementioned $D_{TK}$ logic and the sets of axioms Y and X together with the rules:

NecA. if $D_{TK}\vdash A$ then $D_{TK}[X,Y]\vdash T(A)$

NecB. if $D_{TK}\vdash A$ then $D_{TK}[X,Y]\vdash KNOW_i(A)$

**Theorem 2.**

*$D_{TK}$[X,Y] logics are consistent logics of truth and syntactic modality.*

The following axiom connecting truth and knowledge can be consistently added to the aforementioned logics.

$T(KNOW_i(A))\Leftrightarrow KNOW_i(T(A))$

The predicate $EK_X$ can be defined as a syntactic modality as follows: $EK_X(A) =_{def} E_X(T(A))$

and similarly $CK_X$ as : $CK_X(A) =_{def} C_X(T(A))$

The properties of these two new predicates reflect the properties of the respective modal operators:

E. $EK_X(A)\Leftrightarrow\bigwedge_{i\in X}KNOW_i(A)$

C. $CK_X(A)\Leftrightarrow EK_X^k(A)$ for $k$=1,2,……

IR. If $A \Rightarrow EK_X(A \wedge B)$ then $A \Rightarrow CK_X(B)$

**Theorem 3.**

*The $D_{TK}[X,Y]$ logics together with the above axioms for the $EK_X$ and $CK_X$ predicates, are consistent.*

Consistency for all the above logics can be obtained by using the semantic theory of truth of (Gupta 1982, Herzberger 1982) relativized to possible worlds.

So far all the predicates have taken just a single argument denoting a wff of the language. We now relativize these notions to specific agents in order to truly allow a multi-agent domain. Thus the predicate KNOW is extended so as to take 2 arguments, namely an agent and a wff: KNOW($i$,A) and the predicates EK and CK are extended similarly as follows: EK($X$,A) and CK($X$,A) where $X$ denotes the group of agents that the predicates refer to.

## Common Knowledge and the Surprise Examination

In the preceding sections we mentioned the notion of self-reference and the fact that sometimes it can lead to inconsistencies. We will examine such a situation here, a logical paradox, known as the surprise examination (Montague and Kaplan 1959), which involves self-reference and arises in a multi-agent environment. The situation can be told the following way:

*The teacher tells the class that some time during next week she will give a surprise examination. She will not say on which day for she says it is to be a surprise. Is it possible for the teacher to give a surprise exam?*

A student now can reason as follows. The exam cannot take place on Friday. Because if it hasn't taken place on any of the previous days then on Thursday the student will know that it will take place on Friday. If the student knows the day then it cannot be a surprise, and thus the exam day cannot be Friday. But the exam day cannot be Thursday either. Because if the exam hasn't taken place on any of the previous days and having already excluded Friday, the student thinks that the only possible day left is Thursday. But again if he knows the day then it cannot be a surprise and so the exam day cannot be Thursday. Using the same line of reasoning the student can eliminate the rest of the days as well and come to the conclusion that there is not going to be an exam in the next week.

The paradox has been formalised in a number of ways in the literature and (Sainsbury 1995) contains a very thorough discussion. The teacher's announcement is indeed self-referential. She states on the one hand that there is going to be an exam some time next week and on the other that based on her very announcement (this is where self-reference occurs, the teacher refers to her own announcement) the students will not know the day of the exam, it will come as a surprise. Although the surprise examination has been characterised as a paradox, we believe that other important issues as well, like the means of communication, social relationships and moral obligations which arise, have not been sufficiently explored. As we are going to show in all these issues the concept of common knowledge is crucial and can help us investigate it.

Communication among any members of the group is achieved through announcements and speech acts, which is a way of making someone's plans and intentions known. Recall from the second section that the content of public announcements is considered to be common knowledge among the speaker and the audience after the announcement is made. Since the announcement is made in public this creates a moral obligation as well. The teacher's declaration of her intentions can be interpreted as a promise, or a commitment to the students. We prefer to keep our promises because we do not want to destroy our reputations for not keeping promises. Keeping one's promises is a convention and this in fact is common knowledge among the members of our society.

In addition the surprise examination takes place among the members of a specific organised group, a class. In it the students and the teacher have separate roles each with respective rights and obligations. Furthermore the teacher has a certain kind of authority which is well known and accepted among the other members of the group.

Taking into account the whole setting of the surprise examination and from the discussion above we can positively say that common knowledge arises in this case. A student belongs to a structured group, in which the teacher has an authoritative position which gives her the right to make decisions and the students are obliged to follow them. This is one reason for the student to take the teacher's announcement seriously. Moreover the teacher is considered to be truthful and sincere and this in fact is part of her obligations and commitments towards the class. Furthermore her intentions have taken the form of a promise, a strong social commitment and this is another reason for the student to believe the teacher's announcement. Going even further, if the student tries to model the teacher's reasoning he will be able to see that what the teacher had promised to do is indeed achievable. Therefore it is quite reasonable for the student to conclude that having common knowledge of the teacher's announcement, he has common knowledge of the fact that he is not going to know the exact day of the exam, and this conclusion is supported by the above argumentation. Obviously, in order to fully analyse the surprise examination we would need not only knowledge and common knowledge, but intentions, goals, commitments, obligations and actions. However here we are going to restrict our formal analysis to the use of common knowledge only, since our theory does not fully account for all the issues involved in this complex situation.

The basic idea behind our formalisation of the surprise examination is that the teacher's self-referential announcement does not include the kind of information needed by the class to estimate the exact day of the exam. Therefore knowing the announcement does not imply knowing the day of the exam. These intuitions now take the following form: $A:A_1 \wedge (\text{Know}(C,A) \Rightarrow \neg\, \text{Know}(C,B))$

where A:"I will give you an exam next week and based on the knowledge of this announcement, you will not know the exact day of the exam", $A_1$: "I will give you an exam next week", B: "The exact day of the exam", and $A_1$ and B can be written as first order formulas of our language. For simplicity we do not write the first argument of the predicates that refers to the class. With the introduction of common knowledge, the fact that the class will not know the exact day of the exam becomes common knowledge.

1. A Assumption
2. EK(A) Everyone knows
3. $EK(A_1 \wedge (KNOW(A) \Rightarrow \neg KNOW(B)))$ 2, def. of A
4. $EK(A_1) \wedge EK(KNOW(A) \Rightarrow \neg KNOW(B))$ 3,Thm of $D_{TK}[X,Y]$
5. $EK(KNOW(A) \Rightarrow \neg KNOW(B))$ 4, $\wedge$-elimination
6. $EK(KNOW(A)) \Rightarrow EK(\neg KNOW(B))$ 5,Thm $D_{TK}[X,Y]$
7. $CK(KNOW(A)) \Rightarrow CK(\neg KNOW(B))$ A was public thus it is common knowledge as well as $(\neg KNOW(B))$

We conclude that common knowledge of the announcement implies that it is common knowledge that the class will not know the exact day of the exam. The class is in fact aware of their lack of knowledge of the examination day. This result is quite different from those found in the literature so far (Sainsbury 1995). There is no good reason for rejecting the announcement as a false one since the teacher is considered truthful and has no intention of deceiving the class. It is not logical either to conclude that the class does not know the announcement. The fact that the announcement is made in public puts the students in a very special situation, in which all of them know that the announcement is true and that they are in this situation, the announcement is common knowledge.

## Concluding Remarks.

This paper has presented a method for the formal description of multi-agent systems. Although this work leaves many unanswered questions, we believe it is a first step towards the direction of a first order theory of social agents with self-referential capabilities. A first order language has been presented in which the concepts of truth, knowledge and common knowledge have been formalised based upon the framework for syntactic modalities set up by (Turner 1990). A number of properties have been introduced as axioms of a theory of agency concerning the aforementioned concepts. It is argued that the intuitions captured in this model provide a flexible way of describing agents and the kind of reasoning involved in a multi-agent environment. The framework is quite flexible and for instance, instead of knowledge we can formalise the weaker notion of belief. Our model presents the following advantages when compared with other approaches. Firstly we use first order logic which offers more expressive power than classical modal logic. Furthermore, although the formal counterpart of first order theories, standard modal logics, offer another possible and attractive way for formalising knowledge and belief, they lack the characteristic of self-reference. Once the mechanism for implementing circular reference is added to modal logic, the whole approach runs into problems and modal theories suffer from the same drawbacks as syntactic theories (Perlis 1985). Our approach yields consistent self-referential logics for truth, knowledge and common knowledge which are not too weak to work with in which statements like the "liar" and the "knower" (Montague and Kaplan 1959) can be blocked. We have also formalised the surprise examination based on our intuitions about the teacher's announcement as a self-referential statement, without however changing the meaning of the original statement, and we demonstrated how common knowledge can be employed to investigate it. There are a number of possible avenues for future development of this model. Firstly the notion of time can be incorporated into the model and the relation between time, common knowledge and action can be studied. Secondly the model can be extended to a K(B)DI (Rao and Georgeff 1991) model and enriched by incorporating intentions and desires as new syntactic modalities. However, while the ideas in this paper may be conceptually appealing, considerable work remains to be done to analyse the utility of the approach in more complex situations.

## Acknowledgements

I would like to thank Ray Turner and the reviewers for their helpful comments on an earlier draft of this paper. This research has been supported by GSSF grant 189/96.

## References

Dennet D.C. 1987. *The Intentional Stance*. Cambridge, Mass.: The MIT Press.

Fagin *et al.* 1995. *Reasoning about knowledge*. Cambridge, Mass.: The MIT Press.

Gupta H. 1982. Truth and Paradox. *Journal of Philosophical Logic* 11:1-60.

Herzberger H. 1982. Notes on Naïve Semantics. *Journal of Philosophical Logic* 11:61-102.

Lewis D. 1969. *Convention, A Philosophical Study*. Cambridge, Mass.: Harvard University Press.

Montague R. And Kaplan D. 1960 A Paradox Regained. *Notre Dame Journal of Symbolic Logic* 1:79-90.

Perlis D. 1985. Languages with Self-reference I: Foundations. *Artificial Intelligence* 25:301-322.

Rao A. and Georgeff M. 1991. Modelling Rational Agents within a BDI-Architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*. San Mateo, Calif.: Morgan Kaufmann Publishers.

Sainsbury R.M. 1995. *Paradoxes*. Cambridge, England: Cambridge University Press.

Turner R.1990. *Truth and Modality for Knowledge Representation*. Cambridge, Mass.: The MIT Press.