1 EDITORIAL

Welcome to the first issue of

ΦNEWS

the newsletter for ΦLOG—The Danish Network for Philosophical Logic and Its Applications, sponsored by The Danish National Research Council for the Humanities. ΦNEWS is published twice a year through 2002-2003. The newsletter is freely distributed to the ΦLOG-members and an online copy may be obtained from the ΦLOG homepage. This issue first includes an introduction to ΦLOG, information concerning future activities, an essay motivating the second ΦLOG conference on Selfreference and an extract (in Danish) of the application for ΦLOG funding through The Danish National Research Council for the Humanities. Secondly, this issue includes specific information pertaining to Dimensions in Epistemic Logic, the first conference under the ΦLOG auspices, Roskilde University, May 2-4, 2002. Thus, keep this issue as your conference booklet if you are attending this event.

The next issue of ΦNEWS is scheduled for October 2003. ΦNEWS accepts and publishes contributions in terms of shorter essays, book reviews and other material relevant to philosophical logic and its applications. Please send your contribution to either one of the ΦNEWS editors. Contact information is available on the following page.

April, 2002
Vincent F. Hendricks
Stig Andur Pedersen
Editors

CONTENT

ΦNEWS is published by

ΦLOG
The Danish Network for Philosophical Logic
and Its Applications

Department of Philosophy and Science Studies
Roskilde University, PA6
P. O. Box 260
DK4000 Roskilde, Denmark
Phone: (+45) 4674 2343 Fax: (+45) 4674 3012
Homepage: `http://www.philog.ruc.dk`
ISSN: 1602-1444

Edited by
Vincent F. Hendricks
Stig Andur Pedersen

Volume 1, April 2002

ΦLOG —The Danish Network for Philosophical Logic and Its Applications

is sponsored by The Danish National Research Council for the Humanities for a period of two years starting in January 2002 and ending in December 2003. ΦLOG is an interdisciplinary network in philosophical logic aiming at coordinating research activities and promoting the field of philosophical logic both in Denmark and abroad. ΦLOG is located at the

Department of Philosophy and Science Studies
Roskilde University, PA6
P. O. Box 260
DK4000 Roskilde, Denmark
Phone: (+45) 4674 2343 Fax: (+45) 4674 3012

Membership of ΦLOG is free. Send a letter, fax or email to the network coordinators Vincent F. Hendricks (vincent@ruc.dk) or Stig Andur Pedersen (sap@ruc.dk) to obtain membership. Remember to state your name, position and affiliation and a short description of your interest in philosophical logic.

ΦLOG Background

Today, philosophical logic plays a significant role, not only in philosophy but also in linguistics, computer science, cognitive

psychology and law to mention but a few disciplines. Philosophical logic covers a wide range of logical studies and activities in the humanities and natural science:

- Modal Logic: The study of the moods or modalities with which propositions can be true or false.
- Temporal Logic: The study of time, its structure and properties.
- Epistemic, Doxastic Logic and Knowledge Representation: The study of knowledge, belief and cognitive capabilities, their structure, strength and validity.
- Deontic Logic and Legal Reasoning: The study of the logical structure of ethical judgements and evaluations, and their relations to law.
- Logic Programming: The development of systems for modelling reasoning patterns, intelligent database applications etc.

This list is by no means exhaustive. It is evident however that philosophical logic in general enjoys a wide range of applications and constitutes an interdisciplinary field of research.

ΦLOG Motivation

The network is motivated by a number of reasons:

- Coordination of the Research Activities. Denmark has quite a few internationally highly estimated scholars interested in philosophical logic from very diverse perspectives. These scholars work however mostly in isolation. ΦLOG is an attempt to break the isolation by coordinating (and informing of) the research activities and let the different scholars share information, international contacts, ideas etc. In brief to serve as an engine for interaction.

- The Need for a Qualified Discussion of the Role of Philosophical Logic. With all the different applications of philosophical logic, there is a need for a forum in which researchers with very different interests, specialties and craftsmanship can discuss more general issues related to philosophical logic.

- To Shape and Sharpen the General Interest in Philosophical Logic. Logic and philosophical logic are not generally very popular fields of study in Denmark and abroad. This is perhaps due to the fact that many find it hard to see to what end or to what extent logic can contribute to other areas. Philosophical logic, in contrast to say pure mathematical logic, is intrinsically characterized by a broader perspective. Thus, ΦLOG, given the diversity of the research body, is intended to communicate results to a less specialized public and let the interest in, and general awareness of, philosophical logic enhance.

ΦLOG Activities

- Conferences: The network will be the host of international conferences on philosophical logic 1-2 times a year.

- Seminars and Workshops: The network will host seminars and workshops in Roskilde, Århus, Odense, Kolding and København.

- Ph.D.-courses: The network will arrange international PhD-courses on selected topics within philosophical logic, its philosophical aspects, methods and formal tools.

- Publishing: The network will publish conference proceedings from selected conferences hosted.

ΦLOG Organization

ΦLOG is located the Department of Philosophy and Science Studies at Roskilde University. The coordinators are:

- Vincent F. Hendricks, vincent@ruc.dk
- Stig Andur Pedersen, sap@ruc.dk
- Pelle Guldborg Hansen (secretary), pgh@ruc.dk

The network also has an organizing committee consisting of:

- Torben Bräuner, Computer Science Department, Roskilde University, torben@ruc.dk
- Henning Christiansen, Computer Science Department, Roskilde University, henning@ruc.dk
- Jan Riis Flor, Department of Philosophy, Rhetorics and Education, University of Copenhagen, flor@hum.ku.dk
- Cynthia M. Grund, Department of Philosophy, Odense University, cmgrund@filos.sdu.dk
- Lars Bo Gundersen, Department of Philosophy, Aarhus University, fillg@hum.au.dk
- Per Hasle, Department of Business Communication and Information Science, University of Southern Denmark, hasle@sitkom.sdu.dk
- K. Hvidtfeldt Nielsen, Department of German Philology, Aarhus University, gerkhn@mail.hum.au.dk
- Peter Øhrstrøm, Department of Business Communication and Information Science, University of Southern Denmark, ohrstrom@sitkom.sdu.dk

ΦLOG Homepage

The ΦLOG site on the web is located at

`http://www.philog.ruc.dk`.

Questions, comments and suggestions should be directed to the webmaster Vincent F. Hendricks (vincent@ruc.dk).

SELF-REFERENCE AND LOGIC

▶ Thomas Bolander
Department of Informatics and Mathematical Modelling
Technical University of Denmark
tb@imm.dtu.dk

Self-reference is used to denote any situation in which someone or something refers to itself. Object that refer to themselves are called **self-referential**. Any object that we can think of as referring to something—or that has the ability to refer to something—is potentially self-referential. This covers objects such as sentences, thoughts, computer programs, models, pictures, novels, etc.

The perhaps most famous case of self-reference is the one found in the **Liar sentence**:

> "This sentence is not true".

The Liar sentence is self-referential because of the occurrence of the indexical "this sentence" in the sentence. It is also **paradoxical**.[1] That self-reference can lead to paradoxes is the main reason why so much effort has been put into understanding, modelling, and "taming" self-reference. If a theory allows for self-reference in one way or another it is likely to be inconsistent because self-reference allows us to construct paradoxes, i.e. contradictions, within the theory. This applies, as we will see,

[1] If the sentence is true, what it states must be the case. But it states that it itself is not true. Thus, if it is true, it is not true. On the contrary assumption, if the sentence is not true, then what it states must not be the case and, thus, it is true. Therefore, the sentence is true iff it is not true.

to theories of sets in mathematics, theories of truth in the philosophy of language, and theories of introspection in artificial intelligence, amongst others.

This essay consists of two parts. The first is called "Self-reference" and the second is called "Logic". In the first part we will try to give an account of the situations in which self-reference is likely to occur. These can be divided into situations involving *reflection*, situations involving *universality*, and situations involving *ungroundedness*.[2] In the second part we will turn to a more formal treatment of self-reference, by formalizing a number of the situations involving self-reference as theories of first-order predicate logic. It is shown that Tarski's schema T plays a central role in each of these formalizations.[3] In particular, we show that each of the classical paradoxes of self-reference can be reduced to schema T. This leads us to a discussion of schema T, the problems it gives rise to, and how to circumvent these problems.

The first part of the essay does not require any training in mathematical logic.

Part I: Self-Reference

We start out by taking a closer look at paradoxes related to self-reference.

1 Paradoxes

A paradox is a "seemingly sound piece of reasoning based on seemingly true assumptions, that leads to a contradiction (or other obviously false conclusion)" [Audi, 1995]. A classical example is **Zeno's Paradox** of Achilles and the Tortoise in which we seem to be able to prove that the tortoise can win any race against the much faster Achilles, if only the tortoise is given an arbitrarily small head-start (cf. [Erickson and Fossa, 1998] for

[2] Often cases of self-reference will fit into more than one of these categories.

[3] Tarski's schema T is the set of all first-order logical equivalences

$$T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$

where φ is any sentence and $\ulcorner\varphi\urcorner$ is a term denoting φ.

a detailed description of this paradox). Another classical paradox is the **Liar Paradox**, which is the contradiction derived from the Liar sentence. Among the paradoxes we can distinguish those which are related to self-reference. These are called the **paradoxes of self-reference**. The Liar Paradox is one of these, and below we consider a few of the others.

Grelling's Paradox

A predicate is called **heterological** if it is not true of itself, that is, if it does not itself have the property that it expresses. Thus the predicate "long" is heterological, since it is not itself long (it consists only of four letters), but the predicate "short" is not heterological. The question that leads to the paradox is now:

> Is "heterological" heterological?

It is easy to see that we run into a contradiction independently of whether we answer 'yes' or 'no' to this question.

Grelling's paradox is self-referential, since the definition of the predicate "heterological" refers to *all* predicates, including the predicate "heterological" itself.

Richard's Paradox

Some phrases of the English language denote real numbers. For example, "the ratio between the circumference and diameter of a circle" denotes the number π. Assume that we have given an enumeration of all such phrases (e.g. by putting them into lexicographical order). Now consider the phrase

> "the real number whose nth decimal place is 1 if the nth decimal place of the nth phrase is 2, otherwise 1".

This phrase defines a real number, so it must be among the enumerated phrases, say number k in this enumeration. But, at the same time, by definition, it differs from the number denoted by the kth phrase in the kth decimal place.

Richard's paradox is self-referential, since the defined phrase refers to *all* phrases that define real numbers, including itself.

Berry's Paradox

Berry's Paradox is obtained by considering the phrase

> "the least natural number not specifiable by a phrase containing fewer than 100 symbols".

The contradiction is that that natural number has just been specified using only 87 symbols!

The paradoxes may seem simply like amusing quibbles. We may think of them as nothing more than this when they are part of our imprecise natural language and not part of theories. When the reasoning and assumptions involved in the paradoxes are not attempted to be made completely explicit and precise, we might expect contradictions to be derivable because of this lack of precision. But having a theory—mathematical, philosophical or otherwise—containing a contradiction is of course devastating for the theory. It shows the entire theory to be inconsistent (unsound). The problem is that it turns out that in many of the intuitively correct theories in which some kind of self-reference is taking place, we can actually reconstruct the above paradoxes, and thereby show these theories to be inconsistent. This applies to the naive theories of truth, sets, and introspection as we will later see.

Before we turn to a more thorough study of the situations in which self-reference is to be expected to occur, we put a bit more structure on our notion of self-reference by introducing *reference relations*.

2 Reference Relations

Reference can be thought of as a relation R between a class of referring objects and a class of objects being referred to. R is called a **reference relation**, and it is characterized by the property that

$$(a, b) \in R \quad \text{iff} \quad b \text{ is referred to by } a.$$

The **domain** of R, that is, the set of a's for which there is a b with $(a, b) \in R$, is denoted $\mathrm{dom}(R)$. The **range** of R, that is, the set of b's for which there is an a with $(a, b) \in R$, is denoted $\mathrm{ran}(R)$. The relation R can be depicted as a graph on $\mathrm{dom}(R) \cup \mathrm{ran}(R)$, in which there is an edge from $a \in \mathrm{dom}(R)$ to $b \in \mathrm{ran}(R)$ iff $(a, b) \in R$. If e.g.

$$A = \{\text{"small car"}, \text{"big car"}, \text{"cars"}\}$$

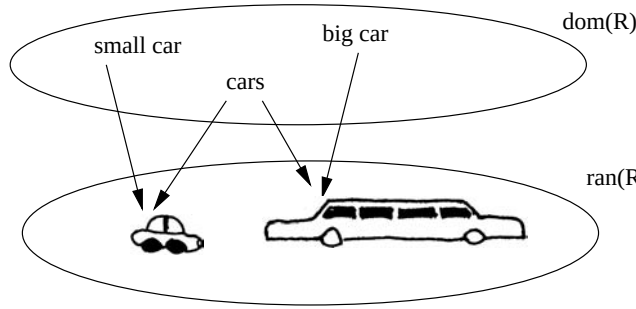


Figure 1 A reference relation.

and

$$B = \left\{ \text{  }, \text{  } \right\}$$

we could have the reference relation depicted on Figure 1. If $\mathrm{dom}(R) \cap \mathrm{ran}(R) = \emptyset$, as above, a referring object will always be isolated from the object it refers to, since these two objects will be members of two distinct and disjoint classes. Self-reference is thus only possible when $\mathrm{dom}(R) \cap \mathrm{ran}(R) \neq \emptyset$.

Let T be the self-referential sentence

"This sentence is true"

(T for **truth teller**). The sentence refers to

(i) the sentence itself

(ii) the "is"-relation

(iii) the concept of truth.

Graphically, this could be represented by the reference relation in Figure 2. Notice the loop at T. The loop means that

$$(T, T) \in R,$$

that is, T is referred to by T, which is exactly the condition for T being self-referential. This leads us to the following definition:

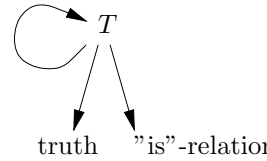


Figure 2 Reference relation for T.

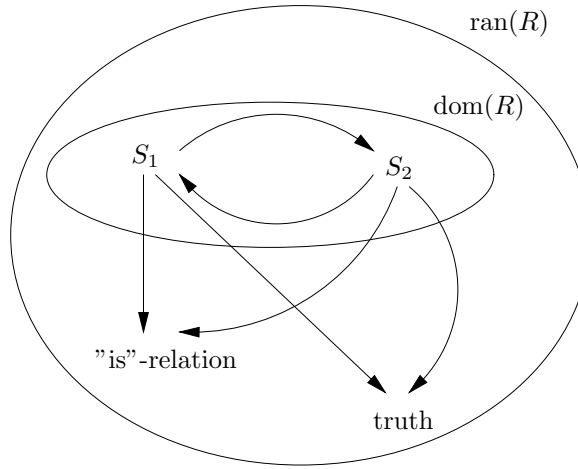


Figure 3 Reference relation for S_1 and S_2.

An object $a \in \mathrm{dom}(R)$ is called **directly self-referential** if there is a loop at a in (the graph of) the reference relation.

Now consider the following two sentences, S_1 and S_2,

S_1 : The sentence S_2 is true.
S_2 : The sentence S_1 is true.

The reference relation for these two sentences become as depicted i n Figure 3. Here the set of referring objects is $\mathrm{dom}(R) = \{S_1, S_2\}$ and the set of objects referred to is

$$\mathrm{ran}(R) = \{S_1, S_2, \text{“is”-relation}, \mathrm{truth}\}\,.$$

Notice that $\mathrm{dom}(R) \cap \mathrm{ran}(R) \neq \emptyset$. None of these sentences are directly self-referential, but S_1 refers to S_2 which in turn refers back to S_1, and vice versa. This gives a *cycle* in the graph consisting of the nodes S_1 and S_2, and the two edges connecting them. We consider both of S_1 and S_2 to be *indirectly self-referential* since each of them refers to itself through the other sentence. Thus we define:

> An object $a \in \mathrm{dom}(R)$ is called **indirectly self-referential** if a is contained in a cycle in (the graph of) the reference relation.

Kripke gives a very nice example of indirect self-reference in [Kripke, 1975]. S_1 is the following statement, made by Jones,

S_1 : Most of Nixon's assertions about Watergate are false.

and S_2 is the following statement, made by Nixon,

S_2 : Everything Jones says about Watergate is true.

The reference relation for this pair of sentences will contain that of Figure 3, i.e. we have again a cycle between S_1 and S_2.

Let us consider a few additional examples of indirect self-reference. In the following, when an object is either directly or indirectly self-referential, we often simply call it **self-referential**.

Naive Set Theory

In naive set theory (as conceived in the early works of Georg Cantor. See e.g. [Cantor, 1932]) the concept of a set can be defined in the following way:

> By a set we understand any collection of mathematical objects (including sets).

We see that the concept of a set is defined in terms of mathematical objects which can themselves be sets. This means that what we have is a self-referential definition of the concept of a set. This self-reference makes the defined concept inconsistent, as we will see from Cantor's Paradox, introduced in Section 2.

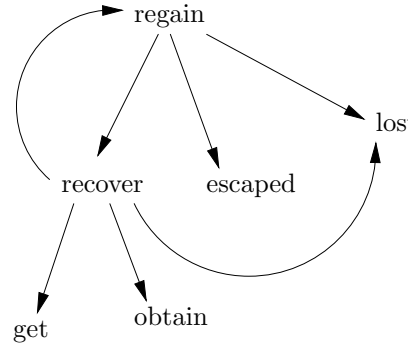


Figure 4 A dictionary reference relation.

Dictionary Reference

In a dictionary, the referring objects are the *definienda*, that is, the expressions or words being defined, and the objects referred to are the *definientia*, that is, the expressions or words that define the definienda. In Webster's 1828 dictionary the word "regain" is defined as:

> *regain* : to *recover*, as what has *escaped* or been *lost*.

At the same time, the word "recover" is defined as:

> *recover* : to *regain*; to *get* or *obtain* that which was *lost*.

Using only the words in italic, the reference relation for the above two dictionary definitions become as depicted in Figure 4. Since the definition of "regain" refers to the word "recover" and the definition of "recover" refers to the word "regain", there is a cycle between these two words in the graph. Each is defined through the other in an indirectly self-referential way. This means that unless we know the meaning of one of these words in advance, the dictionary definition will not be able to give us the full meaning of the other word.

This becomes even worse if we consider an English dictionary of the *entire* English language. Since every word is simply defined in terms of other words, we will not from the dictionary be able to learn the meaning of *any* of the words, unless we know the meaning of some of them in advance. This

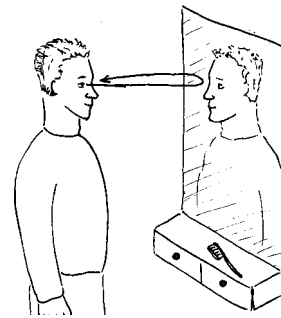


Figure 5 Reflection means "bending back".

makes a dictionary insufficient as a definition of meaning for a language, as noted by Wittgenstein in the so-called *Blue Book* ([Wittgenstein, 1958]). Wittgenstein's way out was to think of a dictionary as supplied with a set of *ostensive definitions*. An **ostensive definition** of a word is a definition "by pointing out" the referent of the word—e.g. to say the word "banjo" while pointing to a banjo.

Wittgenstein's ideas are related to ideas of groundedness of ungroundedness of reference relations, as we will see in Section 5. But before that we will relate self-reference to reflection and universality.

3 Reflection and Self-Reference

Self-reference is often an epiphenomenon of *reflection* of some kind. The word reflection actually means "bending back". We use reflection to denote situations such as: viewing yourself in a mirror; exercising *introspection* (that is, reflecting on yourself and your own thoughts and feelings); having a theory which is contained in its own subject matter; having a picture which contains a picture of itself (Figure 6). Reflection can also be considered as a name for all the situations in which someone or something views itself "from the outside". In the framework of reference relations, we can choose to define:

> A reference relation R is said to have **reflection** if $\mathrm{dom}(R) \subseteq \mathrm{ran}(R)$.

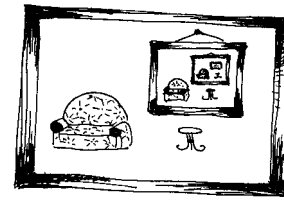


Figure 6 A picture containing itself.

By this definition, a reference relation has reflection iff every referring object is also an object that is referred to. That is, if R is the reference relation of some theory, then that theory can refer (represent, describe) not only objects of the "external world" but also all the objects of the theory itself.

Reflection does not in itself necessarily lead to self-reference, though self-reference often comes together with reflection. We do only have self-reference if we among the elements of $\mathrm{dom}(R)$ can point out an element r which refers to r. Reflection means that every element r of $\mathrm{dom}(R)$ is referred to by another element q of $\mathrm{dom}(R)$, but for all such pairs (q, r) we might have $q \neq r$. In Section 4 we will show, though, that if reflection is combined with universality, then self-reference cannot be avoided.

Below we will consider some important examples of reflection.

Artificial Intelligence

A very explicit form of reflection is involved in the construction of artificial intelligence systems such as for instance robots. Such systems are called **agents**. Reflection enters the picture when we want to allow agents to reflect upon themselves and their own thoughts, beliefs, and plans. Agents that have this ability we call **introspective agents**.

An artificial intelligence agent is most often equipped with some formal language which it uses for representing its experiences and beliefs, and which it uses for reasoning about its environment. That is, such an agent has a model of the world it inhabits which is represented by a set of formal sentences.

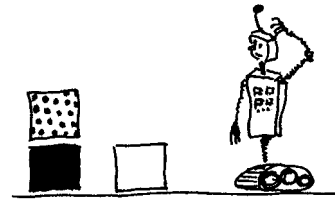


Figure 7 An agent in Blocks World.

Consider an agent situated in a *blocks world*[4] as depicted in Figure 7. The agent's task in this world is to move blocks to obtain some goal configuration (e.g. building a tower consisting of all blocks placed in a specific order). The agent's beliefs about this world could be represented in the agent by formal sentences such as

$$\begin{aligned} &\mathit{on}(\mathit{black\ box}, \mathit{floor}) \\ &\mathit{on}(\mathit{dotted\ box}, \mathit{black\ box}) \\ &\mathit{on}(\mathit{white\ box}, \mathit{floor}) \\ &\mathit{on}(\mathit{agent}, \mathit{floor}). \end{aligned}$$

For the agent to be introspective, though, it should also contain sentences concerning the agent's own beliefs. If the agent believes the sentence

$$\mathit{on}(\mathit{black\ box}, \mathit{floor})$$

to be part of its own model of the world, that could e.g. be represented by the sentence

$$\mathit{agent}(\ulcorner \mathit{on}(\mathit{black\ box}, \mathit{floor}) \urcorner).$$

Now, the referring objects in this situation are obviously the sentences that make up the agent's model of the world. So if R denotes the reference relation of the agent, then $\mathrm{dom}(R)$ consists of all these sentences. The object referred to in the case of a sentence like $\mathit{on}(\mathit{black\ box}, \mathit{floor})$ is the black box on the

[4] "Blocks worlds" are the classical example domains used in artificial intelligence.

floor, while the object referred to in the case of a sentence like $agent(\ulcorner on(black\ box, floor)\urcorner)$ is the sentence $on(black\ box, floor)$. If φ is any sentence then $agent(\ulcorner\varphi\urcorner)$ is a sentence referring to φ. This means that the set of objects referred to, $\mathrm{ran}(R)$, contains every sentence, i.e. we have $\mathrm{dom}(R) \subseteq \mathrm{ran}(R)$. By our definition, this means that R has *reflection*. This reflection—that the agent can refer to any of its own referring objects—turns out to provide a major theoretical obstacle to the construction of introspective agents, as we will see in Section 11.

Philosophy of Language

One of the major problems in the philosophy of language is to give a definition of truth for natural languages. Tarski suggests that every adequate theory of truth should give a predicate "true" satisfying

$$\varphi \text{ is true} \quad \text{iff} \quad \varphi$$

where φ is any sentence. In such a theory of truth we would also have reflection, since the referring objects are sentences, and any sentence φ can be referred to by the sentence

"φ is true".

Reflection is in itself not enough to give self-reference. In both examples above we had reflection but no self-reference, since there were no cycles in the reference relations. The problem is, though, that reflection often comes together with *universality*, and when we have both reflection and universality then self-reference cannot be avoided. Universality is the subject of the following section.

4 Universality and Self-Reference

When we make a statement about all entities in the world, this will necessarily also cover the statement itself. Thus such statements will necessarily be self-referential. We call such statements **universal** (as we call formulas of the form $\forall x\varphi(x)$ in predicate logic). Actually, we will use the term "universal" to denote any statement concerning all entities in the relevant domain of discourse. Correspondingly, in the framework of a reference relation R, we can define:

An object $a \in \mathrm{dom}(R)$ is called **universal** if $(a, b) \in R$ for all $b \in \mathrm{ran}(R)$.

If R is the reference relation of our natural language then the sentence

"All sentences are false" (3.1)

will be universal. The problem about universality is that reflection and universality together necessarily lead to self-reference, and thereby is likely to give rise to paradoxes. To see that reflection and universality together lead to self-reference, assume R has reflection and that $a \in \mathrm{dom}(R)$ is a universal object. Then we have $(a, b) \in R$ for all $b \in \mathrm{ran}(R)$, and since $\mathrm{dom}(R) \subseteq \mathrm{ran}(R)$ we especially get $(a, a) \in R$. That is, we have the following result:

Assume R has reflection and that $a \in \mathrm{dom}(R)$ is universal. Then a is self-referential.

Universality enters the picture in the two examples of reflection previously given if we want the agent to be able to express universal statements about its environment or if we want to be able to apply the truth predicate to sentences that concern all sentences of the language (like e.g. the sentence (3.1)). In such cases self-reference cannot be avoided, and as we will see in the second part of the essay this will allow the paradoxes to surface and produce contradictions in the involved theories.

The problem sketched is not in any way only related to theories of agent introspection and truth. Any theory that is part of its own subject matter has reflection. Thus, if these theories make use of universal statements as well, then these theories contain self-referential statements, and then the paradoxes of self-reference will not be far away. Thus, self-reference is a problem to be taken seriously by any theory that is part of its own subject matter. This applies to theories of cognitive science, psychology, semiotics, mathematics, sociology, system science, cybernetics, computer science.

Note, that each of the paradoxes of self-reference considered in Section 1 involves both reflection and universality, since they all refer to the totality of objects of their own type: the predicate

"heterological" refers to all predicates; the phrase defining a real number in Richard's paradox refers to all phrases defining real numbers; the phrase specifying a natural number in Berry's paradox refers to all phrases specifying natural numbers.

Let us conclude this section by considering another example of a universal object in a reflective setting. In the naive theory of sets (cf. Section 2) we can consider the set U of all sets. U is certainly a universal object, since it refers to all other sets.[5] At the same time, the theory of sets is reflective since for the reference relation R of sets, $\mathrm{dom}(R)$ and $\mathrm{ran}(R)$ are both the class of all sets. Thus U is a self-referential object, and this leads to trouble. Cantor have proved that the cardinality[6] of any set is smaller than the cardinality of the set of subsets of this set. This result is called **Cantor's Theorem**.[7] Let us see what happens if we apply Cantor's Theorem to the set U. First of all, we note that the set of all subsets of U is U itself, since U contains all sets. But then, by Cantor's Theorem, the cardinality of U is smaller than the cardinality of U, which is a contradiction. This contradiction is known as **Cantor's Paradox**. Cantor's Paradox proves that the naive theory of sets is inconsistent.

5 Ungroundedness and Self-Reference

Self-reference often occurs in situations that have an *ungrounded* nature. Given a reference relation R, we can define ungroundedness in the following way:

> An object $a \in \mathrm{dom}(R)$ is called **ungrounded** if there is an infinite path starting at a in the graph of the reference relation R. Otherwise a is called **grounded**.

Note, that if $\mathrm{dom}(R) \cap \mathrm{ran}(R) = \emptyset$, that is, if referring objects and objects being referred to are completely separated, then all elements are grounded.

[5] It is natural to think of the objects being referred to by a set as the elements of the set.

[6] The cardinality of a set is a measure of its size.

[7] It is interesting at this point to note that the argument leading to Cantor's Theorem—a so-called *diagonal argument* (which he was the first to use)—has basically the same structure as Richard's Paradox.

If we take the dictionary example of Section 2, we can give a simple example of ungroundedness. Let R be the reference relation of Webster's 1828 dictionary, that is, let R contain all pairs (a, b) for which b is a word occurring in the definition of a. Since every word of the dictionary refers to at least one other word, every word will be the starting word of an infinite path of R. Here is a finite segment of one of these paths, taken from the 1828 dictionary:

$$\text{regain} \to \text{recover} \to \text{lost} \to \text{mislaid} \to \text{laid} \to \text{position} \to \text{placed} \to \text{fixed} \to \ldots$$

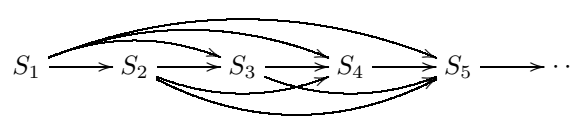
Now the problem that Wittgenstein considered can be stated in the following simple manner: in a dictionary all words are ungrounded. Since there are only finitely many words in the English language, any infinite path of words will contain repetitions. If a word occurs at least twice on the same path, it will be contained in a cycle. Thus, in *any* dictionary of the entire English language there will necessarily be words defined indirectly in terms of themselves. That is, any such dictionary will contain (indirect) self-reference.

Ungroundedness does not always lead to self-reference, but self-reference is very often a byproduct of ungroundedness. So whenever one encounters ungroundedness, one should be very careful to ensure that this ungroundedness does not lead to self-reference and paradoxes.

Actually, as showed by Steven Yablo in [Yablo, 1993], ungroundedness can lead to paradoxes even in cases where we do not have self-reference. **Yablo's Paradox** is obtained by considering an infinite sequence of sentences $S_1, S_2, \ldots$ defined by:

S_1 : All sentences S_i with $i > 1$ are false.
S_2 : All sentences S_i with $i > 2$ are false.
S_3 : All sentences S_i with $i > 3$ are false.
$\vdots$

The reference relation for these sentences looks like this:



As one sees, there is no self-reference involved, but we still get a paradox: Assume S_i is true for some i. Then all S_j for $j > i$ must be false. In particular, S_{i+1} must be false. But since S_j is false for all $j > i + 1$, S_{i+1} must also be true. This is a contradiction. Therefore all S_i must be false. But then S_1 should be true, which is again a contradiction.

It should be noted that even though ungroundedness does not always lead to self-reference, self-reference always leads to ungroundedness: any self-referential object a is contained in a cycle, and we get an infinite path from a by passing through this cycle repeatedly.

6 Vicious and Innocuous Self-Reference

Not all self-reference leads to paradoxes. There is no paradox involved in a self-referential sentence like

"This sentence is true". (3.2)

We can assume either that the sentence is true or that it is false, and neither of the cases will lead into contradiction. But as soon as we introduce a "not" in the sentence, that is, consider the following sentence instead

"This sentence is *not* true" (3.3)

we get a paradox (the Liar Paradox). Self-reference that leads to paradoxes we call **vicious self-reference** and self-reference that does not we call **innocuous self-reference**. It can be shown that self-reference can only be vicious if it involves negation or something equivalent (as the "not" in (3.3)). This means, for instance, that none of the paradoxes of self-reference considered above could be carried through if the occurrence of negation in their central definitions where removed (e.g. if we removed the "not" in the definition of heterological of Grelling's Paradox).

Part II: Logic

We now turn to a more formal treatment of self-reference, by formalizing some of the situations considered in the first part of the essay as theories of first-order predicate logic (henceforth

simply called *first-order theories*). We will assume that all considered first-order theories contain the standard **numerals**:[8]

$$\bar{0}, \bar{1}, \bar{2}, \bar{3}, \ldots$$

We use $\ulcorner \cdot \urcorner$ to range over coding schemes. By a **coding scheme** we understand any injective mapping from sentences into numerals. That is, if φ is a sentence then $\ulcorner \varphi \urcorner$ is the numeral $\bar{n}$ for some natural number n. $\ulcorner \varphi \urcorner$ is a *name* for φ; we call it the **code number** of φ. If $\psi(x)$ is a formula containing x as its only free variable then $\psi(\ulcorner \varphi \urcorner)$ is a sentence expressing that "φ has the property expressed by ψ". In this sense, $\psi(\ulcorner \varphi \urcorner)$ *refers* to φ.

Schema T is, as before, defined as the theory containing each of the equivalences

$$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

where T is a fixed one-place predicate symbol and φ is any sentence.

The aim of this part of the essay is to show that schema T is taking a central position in almost all situations in which we have self-reference. Indeed, schema T can be thought of as a *unifying principle* of all the different occurrences of self-reference.

7 Formalizing Paradoxes

We now try to formalize some of the most famous paradoxes of self-reference to show how these involve schema T. As mentioned, a paradox is a "seemingly sound *piece of reasoning* based on seemingly true *assumptions* that lead to a *contradiction*". Formalizing a paradox means to reconstruct it inside a formal theory (in our case, a first-order theory). This involves finding formal counterparts to each of the elements involved in the informal paradox. The formal counterpart of a "piece of reasoning" is a formal proof and the formal counterpart of an "assumption" is an axiom. Thus the formal counterpart of a piece of reasoning leading to a contradiction will be a formal proof of the inconsistency of the theory in question. Thus:

[8] Actually, any infinite collection of closed terms would do.

A **formalization** of a paradox is a formal proof of the inconsistency of the theory in which the axioms are the formal counterparts of the assumptions of the paradox.

The Liar Paradox

As already mentioned, the Liar Paradox is the contradiction that emerges from trying to determine whether the Liar sentence

"This sentence is false"

is true or false. We will now try to formalize this paradox.

In general, sentences that are directly self-referential can be put in the following form:

$$\text{"This sentence has property } P\text{"}. \tag{3.4}$$

The assumption that characterizes such a sentence is that the term "this sentence" refers to the sentence itself. Another way of stating this assumption is to say that (3.4) should satisfy the following equivalence

$$\begin{aligned} &\text{This sentence has property } P \\ \Leftrightarrow\ &\text{"This sentence has property } P\text{" has property } P, \end{aligned} \tag{3.5}$$

that is, replacing the term "this sentence" by the sentence itself will not change the meaning of the sentence. Formally, this assumption can be expressed as the axiom

$$P(t) \leftrightarrow P(\ulcorner P(t) \urcorner) \tag{3.6}$$

where t is a term having the intended interpretation: "this sentence". This equivalence corresponds to the equivalence (3.5), in that "this sentence" have been replaced by t and the quotes "·" have been replaced by $\ulcorner \cdot \urcorner$.

The Liar Paradox also rests on the assumption that our language has a truth predicate. The formal counterpart of this assumption is that our theory includes schema T. In the Liar sentence, P is the property "not true". Let therefore P in (3.6)

denote the formula $\neg T(x)$. Then, in the theory consisting of schema T and (3.6), we get the following proof:

1.	$\neg T(t) \leftrightarrow \neg T(\ulcorner \neg T(t) \urcorner)$	(3.6) with P being $\neg T$
2.	$T(\ulcorner \neg T(t) \urcorner) \leftrightarrow \neg T(t)$	instance of schema T
3.	$T(\ulcorner \neg T(t) \urcorner) \leftrightarrow \neg T(\ulcorner \neg T(t) \urcorner)$	by 1. and 2.

This proves the theory consisting of (3.6) and schema T to be inconsistent, which is our formalization of the Liar Paradox.

Grelling's Paradox

We will now formalize Grelling's paradox. Recall that Grelling's Paradox is the paradox that emerges when trying to answer whether "heterological" is heterological. The formal counterpart of a predicate is a formula. A formula $\varphi(x)$ is then *heterological* if it is "not true of itself", that is, if

$$\neg T(\ulcorner \varphi(\ulcorner \varphi \urcorner) \urcorner)$$

holds, where T is a truth predicate. So to formalize Grelling's paradox we again need to have schema T among our axioms. We also need axioms that allow us to apply a formula to itself (that is, the code of itself). To obtain this, we introduce a function symbol *app* and axioms

$$app(\ulcorner \varphi(x_1) \urcorner, \tau) = \ulcorner \varphi(\tau) \urcorner \tag{3.7}$$

for all formulas φ and all terms τ. These axioms ensure us that $app(\ulcorner \varphi(x_1) \urcorner, \tau)$ denotes the result of "applying" $\varphi(x_1)$ to τ (that is, instantiating $\varphi(x_1)$ with τ). Now we can formalize the predicate "heterological" as the formula $het(x_1)$ given by

$$het(x_1) =_{df} \neg T(app(x_1, x_1)).$$

To obtain the contradiction we should ask whether $het(\ulcorner het(x_1) \urcorner)$ holds or not. We get the following proof:

1.	$het(\ulcorner het(x_1) \urcorner) \leftrightarrow \neg T(app(\ulcorner het(x_1) \urcorner, \ulcorner het(x_1) \urcorner))$	by def. of $het(x_1)$
2.	$het(\ulcorner het(x_1) \urcorner) \leftrightarrow \neg T(\ulcorner het(\ulcorner het(x_1) \urcorner) \urcorner)$	by 1. and (3.7)
3.	$het(\ulcorner het(x_1) \urcorner) \leftrightarrow T(\ulcorner het(\ulcorner het(x_1) \urcorner) \urcorner)$	instance of schema T
4.	$\neg T(\ulcorner het(\ulcorner het(x_1) \urcorner) \urcorner) \leftrightarrow T(\ulcorner het(\ulcorner het(x_1) \urcorner) \urcorner)$	by 2. and 3.

This proves the theory consisting of (3.7) and schema T to be inconsistent, which is our formalization of Grelling's paradox.

Richard's Paradox is formalized in much the same way as Grelling's Paradox, though the formalization becomes slightly more technical. For these reasons we choose to leave out a formalization of Richard's Paradox in this essay.

Berry's Paradox

Obviously, to formalize Berry's Paradox, we need axioms formalizing a reasonable part of arithmetic. Apart from this we only need a formal counterpart of the notion of specifiability (the formal counterpart of a "phrase" naturally being a formula). We can use the same trick as we did in the previous examples. A formula $\varphi(x)$ specifies the number n iff $\varphi(m)$ holds exactly when $m = n$. If we want to define a formula $spec(x, y)$ such that $spec(\ulcorner\varphi(x)\urcorner, n)$ holds precisely when $\varphi(x)$ specifies n, then it should look like

$$spec(x, y) =_{df} \forall z\,(z = y \leftrightarrow T(app(x, z)))$$

where T and app are defined as before. We will not go further into the details of formalizing this paradox, but refer to [Boolos, 1989] in which this is carried out. We just note that again schema T is central to the formalization. The notion of specifiability could not have been formalized without schema T or something equivalent.

We have now shown how to formalize several of the most famous paradoxes of self-reference, and, as we have seen, these formalized paradoxes all turn out to be reducible to schema T. That is, all these paradoxes have a common core which is schema T. What we can conclude is that:

(i) That schema T can be extracted from all these paradoxes helps us see the close formal relationship between the paradoxes of self-reference.

(ii) That all these paradoxes can be extracted from schema T helps us to see the importance of schema T in understanding the paradoxes of self-reference, and in understanding self-reference in general.

Below we consider some examples of occurrences of schema T in the philosophy of language, mathematics, and artificial intelligence.

8 The Naive Theory of Truth

As mentioned, Tarski thought of his schema T as describing the principle that any theory of truth should satisfy. The first-order theory consisting only of schema T is consistent. But for schema T to be a sensible principle of truth we must expect it to be consistent also when added to any consistent, "realistic" first-order theory. It should be a principle of truth working no matter which domain of discourse we would like to apply truth to. But, unfortunately, because of self-reference it is not so. In the formalizations of the paradoxes above we have seen several examples showing that schema T becomes inconsistent when added to even quite weak and harmless axioms (at least harmless when these axioms are taken by themselves or together with standard theories for arithmetic, set theory, or the like). In fact, it can easily be shown that all of the axioms assumed above in addition to schema T are *interpretable* in Peano Arithmetic, that is, they can all be translated into equivalent axioms of Peano Arithmetic.[9] This gives us the famous **Tarski's Theorem**:

> Peano Arithmetic extended with schema T is inconsistent.

Note the interesting fact that *any* of the above paradoxes can be used to prove Tarski's Theorem—one just needs to show that the axioms of the formalized paradox are interpretable in PA (Peano Arithmetic). This shows that the contradiction derivable from the formalized paradox can be carried through in PA + schema T.

That schema T becomes inconsistent when standard arithmetic is added is a very serious drawback for the theory of

[9] For a precise definition of "interpretable in" we refer to [Mendelson, 1997] or a similar introduction to mathematical logic. At this point it is enough to note that when an axiom A is interpretable in a theory K it means that any proof in $K + A$ can be translated into a corresponding proof in K. It should be noted that to prove the interpretability we need to choose our coding scheme $\ulcorner \cdot \urcorner$ with care.

truth expressed through schema T. It gives rise to an important problem of how we can restrict schema T to regain the essential consistency. This is the question that we take up in Section 12.

But let us first consider some more examples of situations in which schema T turns up, which makes the reasons to find consistent ways to restrict schema T even more urgent.

9 Gödel's Incompleteness Theorem

We now show how schema T is related to Gödel's famous First Incompleteness Theorem.

A version of the Incompleteness Theorem states that

If PA is ω-consistent[10] then it is incomplete.[11]

To prove this, we can show that the assumption that PA is both ω-consistent and complete leads to a contradiction. On the basis of the formalizations of paradoxes that we have been considering, we see that this could be proved by showing that if PA were both ω-consistent and complete then some paradox would be formalizable in PA. This was, roughly, Gödel's idea.[12] He constructed a formula Bew (for "Beweis") in his theory satisfying, for all φ and all n,

$$\vdash Bew(\bar{n}, \ulcorner\varphi\urcorner) \quad \Leftrightarrow \quad n \text{ denotes a proof of } \varphi. \tag{3.8}$$

Assuming the theory to be ω-consistent and complete we can prove that

$$\vdash \exists x Bew(x, \ulcorner\varphi\urcorner) \quad \Leftrightarrow \quad \vdash \varphi$$

for every sentence φ. The proof runs like this: First we prove the implication from left to right. If $\vdash \exists x Bew(x, \ulcorner\varphi\urcorner)$ then there is some n such that $\nvdash \neg Bew(\bar{n}, \ulcorner\varphi\urcorner)$, by ω-consistency. By completeness we get $\vdash Bew(\bar{n}, \ulcorner\varphi\urcorner)$ for this n. By (3.8) above we get that n denotes a proof of φ. That is, φ is provable, so we have $\vdash \varphi$. To prove the implication from right to left, note

[10] A theory is called **ω-consistent** if, for every formula $\varphi(x)$ containing x as its only free variable, if $\vdash \neg\varphi(\bar{n})$ for every natural number number n, then it is not the case that $\vdash \exists x\varphi(x)$.

[11] A theory is incomplete if it contains a formula which can neither be proved nor disproved.

[12] Though he considered a different formal theory, P.

that if $\vdash \varphi$ then there must be an n such that $\vdash Bew(\bar{n}, \ulcorner\varphi\urcorner)$, by (3.8). From this we get $\vdash \exists x Bew(x, \ulcorner\varphi\urcorner)$, as required. This concludes the proof.

Now, when we have

$$\vdash \exists x Bew(x, \ulcorner\varphi\urcorner) \quad \Leftrightarrow \quad \vdash \varphi$$

in a complete theory, we must also have

$$\vdash \exists x Bew(x, \ulcorner\varphi\urcorner) \leftrightarrow \varphi.$$

If we let the formula $\exists x Bew(x, \ulcorner\varphi\urcorner)$ be abbreviated by $T(\ulcorner\varphi\urcorner)$ then these equivalences read

$$\vdash T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$

which is schema T!

That is, if we assume PA (or a related theory) to be ω-consistent and complete then schema T turns out to be interpretable in it. Now, Tarski's Theorem shows that there exists no such consistent theory. This gives us a proof of Gödel's Incompleteness Theorem. Furthermore, in the same way that one could use any of the paradoxes of self-reference to prove Tarski's Theorem, one can use ones favorite paradox of self-reference to prove Gödel's Theorem.

To summarize the process: first you assume your theory to be both ω-consistent and complete. Then you show that this makes schema T interpretable in the theory. Having schema T means that you can choose any paradox of self-reference and formalize it in the theory. The formalized paradox produces a contradiction in the theory, and thus shows that the theory cannot be both ω-consistent and complete.

Gödel himself actually had a footnote in his 1931 article, in which he proved the Incompleteness Theorem ([Gödel, 1931]), saying that any paradox of self-reference[13] could be used to prove the Incompleteness Theorem.

The reason that we have a result such as Gödel's Incompleteness Theorem is closely related to *reflection*. What Gödel ingeniously discovered was that formal theories can be reflected inside themselves, since numerals can be used to refer to formulas through the use of a coding scheme, $\ulcorner\cdot\urcorner$, and by means

[13] He used the term "epistemic" about these paradoxes.

of these codes provability can be restated inside the theories as arithmetical properties.

10 Axiomatic Set Theory

Schema T also plays a central role in axiomatic set theory. By the **full abstraction principle** we understand the set of formulas on the form

$$\forall x \, (x \in \{y \mid \varphi(y)\} \leftrightarrow \varphi(x)) \, ^{14}$$

where φ is any formula. When Gottlob Frege tried to give a foundation for mathematics (set theory) through his works "Die Grundlagen der Arithmetik (1884)" and "Grundgesetze der Arithmetik (1893,1903)", the full abstraction principle were among his axioms. But in 1902 his system was shown to be inconsistent by Bertrand Russell. Russell constructed a paradox of self-reference which was formalizable within Frege's system. **Russell's Paradox** runs like this:

> Let M be the set of all sets that are not members of themselves. Is M a member of itself or not?

From each answer to this question the opposite follows. Notice the similarity between this paradox and Grelling's paradox considered in Section 1. **Russell's Paradox** can be formalized in any system containing the full abstraction principle. We let $M = \{y \mid y \notin y\}$, that is, $M = \{y \mid \varphi(y)\}$ where $\varphi(y) = y \notin y$. The abstraction principle instantiated by the formula φ now becomes

$$\forall x \, (x \in \{y \mid y \notin y\} \leftrightarrow x \notin x) \, .$$

Letting $x = \{y \mid y \notin y\}$, we get

$$\{y \mid y \notin y\} \in \{y \mid y \notin y\} \leftrightarrow \{y \mid y \notin y\} \notin \{y \mid y \notin y\}$$

which is a contradiction. Thus Frege's system, or indeed any system containing the full abstraction principle, is inconsistent.

The discovery of this inconsistency lead to extensive research in how the full abstraction principle could be restricted to regain consistency.

[14] The formula can be read: "for all sets x, x is in the set of y's for which $\varphi(y)$ holds if and and only if $\varphi(x)$ holds".

Actually, as we will now show, every instance of schema T can be interpreted in the corresponding instance of the abstraction principle. This means that if we can prove that a set of instances of schema T is inconsistent, then we have also proven that the corresponding set of instances of the abstraction principle is inconsistent. In other words, proving consistency results about restricted versions of schema T will also give corresponding consistency results about restricted versions of the abstraction principle.

The result is the following:

> Every instance of schema T:
>
> $$T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$
>
> can be interpreted in the corresponding instance of the abstraction principle:
>
> $$\forall x\,(x \in \{y \mid \varphi\} \leftrightarrow \varphi)\,.$$

The proof is quite simple. If we have got a theory containing

$$\forall x\,(x \in \{y \mid \varphi\} \leftrightarrow \varphi)$$

then $T(\ulcorner\varphi\urcorner)$ can be interpreted in it by the following *extension by definitions*:[15]

$$\ulcorner\varphi\urcorner = \{y \mid \varphi\}$$
$$T(x) \leftrightarrow \bar{0} \in x.$$

Since we have

$$\forall x\,(x \in \{y \mid \varphi\} \leftrightarrow \varphi)$$

we get in particular

$$(\bar{0} \in \{y \mid \varphi\} \leftrightarrow \varphi)$$

which is the same as

$$T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi,$$

[15] We refer again to [Mendelson, 1997] for a definition of the concept of "extension by definitions".

using the definitions of $\ulcorner \cdot \urcorner$ and T. This proves $T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ to be interpretable in $\forall x\,(x \in \{y \mid \varphi\} \leftrightarrow \varphi)$.

11 Agent Introspection

We now turn to our last example of an occurrence of schema T in a situation dealing with self-reference. We consider again the problem of constructing introspective agents, as introduced in Section 3. Since the agent's model of the world is supposed to consist of a set of sentences, we can think of this model as being a formal theory K. This could be a theory in any kind of formal language, but at this point we will assume that it is a theory in a first-order language. Then, for the agent to believe that e.g. the black box is on the floor would correspond to having

$$K \vdash \mathit{on}(\mathit{black\ box}, \mathit{floor}). \tag{3.9}$$

If the agent has introspection, it also has beliefs about its own model of the world. If it believes the sentence in (3.9) to be contained in its own model of the world we would have

$$K \vdash \mathit{agent}\,(\ulcorner \mathit{on}(\mathit{black\ box}, \mathit{floor})\urcorner)\,.$$

Now, if we assume that all of the agent's beliefs about itself to be correct, we should have

$$K \vdash \mathit{agent}(\ulcorner \varphi \urcorner) \quad \Leftrightarrow \quad K \vdash \varphi$$

for all sentences φ. Of course, not all of an agent's beliefs about itself will necessarily always be correct. But even so, the agent might believe this to be the case; and that would correspond to having

$$K \vdash \mathit{agent}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

for all sentences φ. Using "T" instead of "*agent*" this gives us, once again, schema T!

That is, if an agent has introspection and believes this introspection to be correct, then it will necessarily contain schema T in its model of the world. As we know from Tarski's Theorem and our formalized paradoxes this is very difficult to obtain without running into contradictions. At least, it is extremely sensitive to what other axioms we have in K. This is a major drawback in the design of introspective agents.

We have to expect that any kind of axioms could be in K, depending on the environment of the agent and its beliefs about it. The set of axioms of K could even change over time due to changes in the environment. If K includes schema T it means that the agent could suddenly become inconsistent as a consequence of changes in the external world. This seems to prove that it is not possible for an introspective agent consistently to obtain and retain the belief that its introspection is correct.

This conclusion appears very counterintuitive, but again it has to do with the paradoxes of self-reference. If the agent has introspection, and believes this introspection to be correct, it can construct paradoxes of self-reference concerning its own beliefs, and these paradoxes make the agent inconsistent.

The problem is now to find ways to treat agent introspection such that this introspection will not lead into inconsistency. It seems that we have two possibilities: either to ensure that the agent will not be able to make self-referential statements (which would be a restriction on its introspective abilities) or to restrict its logical abilities such that self-referential statements could be assumed consistently. Such restrictions are the subject of the following section.

12 Taming Self-Reference

We have now seen that schema T occurs as the natural principle in a large number of situations of very different kinds. Schema T is the underlying principle in the naive theories of truth, sets, and agent introspection. But unfortunately, schema T is also the underlying principle in the paradoxes of self-reference, which means that most of the theories we are interested in become inconsistent when schema T is added. Since the inconsistency of schema T is a consequence of the presence of self-referential sentences, there seems to be two possible ways to get rid of the problem: ban self-referential sentences in our language or weaken the underlying logic so that these sentences will do no harm. That is, the different ways to restrict schema T in order to ensure consistency seems to divide into the following two major categories:

(i) Cutting away the problematic part (i.e. getting rid of the viciously self-referential sentences).

(ii) Making the problematic part unproblematic (i.e. ensure that self-reference does not lead to disaster).

Cutting Away the Problematic Part

Cutting away the problematic part means to restrict the set of instances of schema T such that the viciously self-referential sentences are excluded from entering the schema. By the **T-scheme** over M, where M is a set of sentences, we understand the following set of equivalences:

$$T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi, \text{ for all } \varphi \in M.$$

If M does not contain sentences that are viciously self-referential, it can be proven that the T-scheme over M can consistently be added to *any* consistent theory.[16] This is because banning the viciously self-referential sentences from schema T makes it impossible to reconstruct the paradoxes of self-reference within the theory.

One very coarse way of disallowing self-reference was proposed by Tarski himself ([Tarski, 1956]): M should not be allowed to contain any sentence in which the predicate symbol T occurs. Note that this will ensure that none of the proofs of the formalized paradoxes considered in Section 7 can be carried through. This restriction is sufficient to reestablish consistency, but it is at the expense of a substantial loss of the expressive power of schema T. It means that iterated truth like in

> "It is true that it is not true that n is a prime number"

that formally looks like this

$$T\,(\ulcorner\neg T(prime(\bar{n}))\urcorner)$$

will not be treated correctly by the restricted T-scheme.

Several less coarse solutions have been proposed in the literature since Tarski. First of all, one notes that not all self-reference is vicious, so we can allow self-referential sentences

[16] That is, can consistently be added to any consistent theory that does not in advance contain axioms for the T predicate.

in M as long as they are not vicious. As mentioned, for self-reference to be vicious, it needs to involve negation. A sentence in which the predicate symbol T is not in the scope of negation ($\neg$) is called a **positive sentence**. Positive sentences can be self-referential, but only of the innocuous kind. Donald Perlis and Solomon Fefermann showed independently ([Perlis, 1985], [Feferman, 1984]) that the T-schema over a set of positive sentences can consistently be added to any consistent theory.

Another way to exclude viciously self-referential sentences is to make restrictions on universality. As we saw in Section 4, reflection only necessarily leads to self-reference when it is combined with universality. Refraining from having universal sentences about truth like e.g.

"All sentences are true"

in M we can again obtain a consistent, restricted T-scheme. More precisely, in [Bolander, 2002] it is shown that if none of the sentences of M contain $T(x)$ as a sub-formula with x quantified, then the T-scheme over M can consistently be added to any consistent theory.

Finally, the method of restricting negation and the method of restricting universality can be combined to get an even stronger T-scheme. M can consistently be allowed to contain any sentence in which $T(x)$ does not occur in the scope of negation (see [Bolander, 2002]).

Making the Problematic Part Unproblematic

Another way of ensuring consistency is to stick with self-reference (i.e. all instances of schema T) but to make sure that self-reference does not get the chance to become paradoxical. Such solutions seem again to divide into two categories:

(i) Restricting the form of schema T.

(ii) Restricting the underlying logic.

Below we consider each of these methods.

Restricting the Form of Schema T

Instead of having bi-implications

$$T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \tag{3.10}$$

in some cases it is sufficient to have e.g. the following implications

$$T(\ulcorner\varphi\urcorner) \rightarrow \varphi \quad \text{and} \quad T(\ulcorner\varphi\urcorner) \rightarrow T(\ulcorner T(\ulcorner\varphi\urcorner)\urcorner).$$

Some of these *restrictions on the form of schema T* will form consistent extensions to any consistent theory, even if we do not restrict the set of sentences that these schemas are instantiated with. Results of this type can be found in e.g. [Montague, 1963], [Thomason, 1980], and [McGee, 1985]. Another possibility is to use a weak equivalence operator in (3.10) instead of the classical bi-implication operator $\leftrightarrow$. A result concerning such a weak equivalence operator can be found in [Feferman, 1984].

Restricting the underlying logic

Theories containing schema T become inconsistent because in them we can construct self-referential sentences that turn out to be true iff they are false. If we change the underlying logic such that sentences are allowed either to be nor true nor false, or both true and false, the self-referential sentences will no longer be able to prove the theories to be inconsistent. Kripke considers in [Kripke, 1975] the possibility of allowing sentences to have no truth-value, that is, to be neither true nor false. His trick is then to only assign truth-values to the *grounded sentences* of the language (cf. Section 5). By this, he ensures that no self-referential sentence will be given a truth-value (since every self-referential sentence is ungrounded). This corresponds to the fact that in a dictionary, as considered in Section 5, we can only, from the dictionary alone, assign meaning to the grounded words. The ungrounded words, among these the self-referentially defined ones, will be "undecided" (not be assigned any meaning). Kripke's theory can be used to construct formal systems in which we consistently have schema T, but in which the underlying logic is restricted (we cannot have classical negation, for instance, since this requires every sentence to be either true or false). Graham Priest (in [Priest, 1989] and others) proposes that we should allow sentences to be *both* true and false, because this is, in a sense, what paradoxical self-referential sentences are.

13 Conclusion

The paradoxes of self-reference still have no final solution that is generally agreed upon. This makes them, in a sense, *genuine paradoxes*. The presence of a paradox is always a symptom that some part of our fundamental understanding of a subject is crucially flawed. In Zeno's Paradox it was the understanding of infinity that was deficient. In the paradoxes of self-reference it seems that what we do not yet have a proper understanding of is the fundamental relation between something that *refers* (or *represents*) and something that *is referred to* (or *represented*) when these two can not be completely separated. As long as this relationship is not entirely grasped we will probably not get to a full understanding of the paradoxes of self-reference and their consequences for the theories of truth, sets, agent introspection, etc.

In Zeno's Paradox it was not an explicitly stated assumption that later proved to be defective. In the paradox it was implicitly assumed that "infinitely many things can not happen in finite time", but it was not until the development of the mathematical calculus that this assumption could be made explicit and rejected. In the case of Zeno's Paradox it was thus not simply a question of finding the failing assumption involved in the paradox. It was rather a question of discovering a new dimension of the world that had hitherto been hidden to the human eye. A similar thing might very well be the case for the paradoxes of self-reference. The right solution (assuming there is one) to the paradoxes is not to remove or restrict any of our explicit assumptions (that is, restrict schema T, underlying logic, or similar), but to discover a new dimension of the problem that will in the end give more, not fewer, axioms in some kind of extended logic. This new dimension is then expected to make explicit some assumptions about the general relations between referring objects and objects referred to; assumptions that a now invisible to us.[17]

[17]It has been proposed to solve the paradoxes of self-reference by extending logic to include *contexts*. Then a paradox such as the Liar will be resolved by the fact that the context before the Liar sentence is uttered is different from the context after it has been uttered. Such a solution seems to be of the kind I propose here, since the logic is extended (with contexts) rather than restricted. But, it should be noted that if we are allowed freely

Finally: What would be a suitable concluding remark in an essay like this?[18]

to refer to contexts then the Liar sentence can be strengthened to give a paradox even in this theory.

[18]Answer: A self-referential question which is its own answer.

Bibliography

[Audi, 1995] Audi, R., editor (1995). *The Cambridge Dictionary of Philosophy*. Cambridge University Press.

[Bartlett, 1992] Bartlett, S. J., editor (1992). *Reflexivity—A Source-Book in Self-Reference*. North-Holland, Amsterdam.

[Bolander, 2002] Bolander, T. (2002). Restricted truth predicates in first-order logic. *(submitted for publication)*. To be presented at LOGICA 2002.

[Boolos, 1989] Boolos, G. (1989). A new proof of the Gödel Incompleteness Theorem. *Notices of the American Mathematical Society*, 36:388–390. Reprinted in [Boolos, 1998].

[Boolos, 1998] Boolos, G. (1998). *Logic, Logic, and Logic*. Harvard University Press.

[Cantor, 1932] Cantor, G. (1932). *Gesammelte Abhandlungen*. Springer Verlag.

[Erickson and Fossa, 1998] Erickson, G. W. and Fossa, J. A. (1998). *Dictionary of paradox*. University Press of America.

[Feferman, 1984] Feferman, S. (1984). Toward useful type-free theories. I. *The Journal of Symbolic Logic*, 49(1):75–111. Reprinted in [Martin, 1984].

[Gaifman, 1992] Gaifman, H. (1992). Pointers to truth. *Journal of Philosophy*, 89(5):223–261.

[Gilmore, 1974] Gilmore, P. C. (1974). The consistency of partial set theory without extensionality. In *Axiomatic set theory*

(Proc. Sympos. Pure Math., Vol. XIII, Part II, Univ. California, Los Angeles, Calif., 1967), pages 147–153. Amer. Math. Soc.

[Gödel, 1931] Gödel, K. (1931). Über formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38:173–198. Reprinted in [Gödel, 1986].

[Gödel, 1986] Gödel, K. (1986). *Collected works. Vol. I.* Oxford University Press. Publications 1929–1936, Edited and with a preface by Solomon Feferman.

[Hatcher, 1982] Hatcher, W. S. (1982). *The Logical Foundations of Mathematics*. Pergamon Press.

[Kripke, 1975] Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy*, 72:690–716. Reprinted in [Martin, 1984].

[Martin, 1984] Martin, R. L., editor (1984). *Recent Essays on the Liar Paradox*. Oxford University Press.

[McGee, 1985] McGee, V. (1985). How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic*, 14(4):399–410.

[McGee, 1992] McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic*, 21(3):235–241.

[Mendelson, 1997] Mendelson, E. (1997). *Introduction to Mathematical Logic*. Chapman & Hall, 4 edition.

[Montague, 1963] Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflection principles and finite axiomatizability. *Acta Philosophica Fennica*, 16:153–166.

[Perlis, 1985] Perlis, D. (1985). Languages with self-reference I. *Artificial Intelligence*, 25:301–322. Reprinted in [Bartlett, 1992].

[Perlis and Subrahmanian, 1994] Perlis, D. and Subrahmanian, V. S. (1994). Meta-languages, reflection principles and self-reference. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 2, pages 323–358. Oxford University Press.

[Priest, 1989] Priest, G. (1989). Reasoning about truth. *Artificial Intelligence*, 39(2):231–244.

[Priest, 1994] Priest, G. (1994). The structure of the paradoxes of self-reference. *Mind*, 103(409):25–34.

[Tarski, 1956] Tarski, A. (1956). The concept of truth in formalized languages. In *Logic, semantics, metamathematics*. Hackett Publishing Co., Indianapolis, IN. Papers from 1923 to 1938.

[Thomason, 1980] Thomason, R. H. (1980). A note on syntactical treatments of modality. *Synthese*, 44(3):391–395.

[Wittgenstein, 1958] Wittgenstein, L. (1958). *Blue and Brown Books*. Blackwell.

[Yablo, 1993] Yablo, S. (1993). Paradox without self-reference. *Analysis*, 53(4):251–252.

4 ΦLOG ACTIVITIES, 2002-2003

▶ CONFERENCE —Roskilde University, October 31–November 2, 2002

SELF-REFERENCE

▶ ABSTRACT

Self-reference is used to denote any situation in which someone or something refers to itself. Self-reference is an important issue in philosophy, mathematics and computer science amongst others. In the philosophy of language the naive theory of truth has been challenged by the Liar Paradox. The Liar Paradox is the contradiction that emerges from trying to determine whether the sentence

'This sentence is false'

is true or false. The sentence is obviously self-referential in that it claims itself to be false. In mathematics the naive concept of set has been challenged by Russell's paradox. Russell's paradox is the contradiction that emerges from trying to determine whether the sentence

'Is the set of all sets that are not
members of themselves an element of itself?'

is true or false. This sentence, as well, involves self-reference, though maybe not in a way as obvious as the Liar sentence. In computer science one of the important problems is the question of how to implement introspection (self-reflection) in artificial intelligence agents. Through introspection an agent is able to refer to itself. On the naive account of agent introspection this again leads to a paradox of self-reference, e.g. in the form of the Knower's Paradox:

'I know that what I say now is not true.'

In the light of these paradoxes the naive theories have to be abandoned and several new, consistent theories are introduced instead. In these theories the paradoxes are avoided either by blocking self-reference altogether or by finding consistent ways to treat self-reference. The blocking strategy will most often result in theories that are limited in important ways. Thus, to construct powerful, consistent theories one has to get to a deeper theoretical understanding of self-reference and of how to live consistently with it. It turns out that all three paradoxes above are structurally similar. This implies that coming to an understanding of the basic structure involved in self-reference and theoretically investigate how to tame it has promising perspectives for all three fields of research.

Self-reference is not in any way restricted to occur only in the theories considered above. Actually, any theory that could be considered to be part of its own subject matter has some degree of self-referentiality. This applies to many theories of language, economy, sociology, psychology, etc. With respect to these theories an understanding of self-reference is essential to avoid performing unsound self-referential reasoning as in the paradoxes above. The aim of the conference is to bring together researchers in the fields of philosophy, mathematics, and computer science to present theories of and related to self-reference – especially with respect to theories that explain and resolve the above paradoxes and thereby advance new theories for the involved fields.

▶ SPEAKERS

- Andrea Cantini, Deparment of Mathematics, University of Florence, Italy
- Anil Gupta, Department of Mathematics, University of Indiana and Department of Philosophy, University of Pittsburgh, USA
- Melvin Fitting, Department of Mathematics and Computer Science, Lehman College, New York, USA
- Vann McGee, Department of Philosophy, MIT, USA
- Donald Perlis, Department of Computer Science, University of Maryland, USA

- Graham Priest, Department of Philosophy, University of Melbourne and Department of Philosophy, University of St. Andrews, Australia / Scotland
- Raymond M. Smullyan, Department of Mathematics and Department of Philosophy, University of Indiana, USA
- Stephen Yablo, Department of Linguistics & Philosophy, MIT, USA

▶ Conference: Temporal Logic. Scheduled for spring 2003, University of Southern Denmark, Denmark.

▶ Summer School: Philosophical Logic. Scheduled for summer 2003, Roskilde University, Denmark.

▶ Conference: FOL75. Scheduled for fall 2003, to be held in Berlin in association with Humboldt University, Berlin, Germany.

5 COPENHAGEN CONQUERED BY LOGICIANS THIS SUMMER!

▶ Henning Christiansen

Computer Science Department
Roskilde University
henning@ruc.dk

Copenhagen will be unusual the two last weeks of July when hordes of strange people invade the city. Tourists and inhabitants will be confused by weirdos speaking in all different tongues including English with various accents about abstracts things that no one can understand (often not even themselves). These subjects will be spreading their own very special kind of graffiti on every blank surface, the city's restaurants will find their napkins and table cloths completely decorated by strange 'tags'. Waiters, bus drivers, shop assistants, everyone on the street will be very confused when addressed with nested negations (even different sorts of negations), complicated quantifier structures and higher-order questions.

In early August things will gradually turn into normal as these strange individuals after

$$0 \vee 1 \vee 2 \vee \ldots$$

days of rest in the city spread out in the world to their lonely residences speculating about their strange things, and when and where they can meet so many people of their own kind again.

More than 1000 logicians and computer scientists is expected to participate in

FLoC'02, Federated Logic Conferences
July 20–August 1, 2002

H.C.Ørsted Instituttet, University of Copenhagen
http://floc02.diku.dk

This arrangement comprises 7 international conferences and 33 workshops in the fields between mathematical and philosophical logics and computer science.

I, myself, the author of this article, is chairing the *International Conference on Logic Programming* which is the eighteenth of its kind, dating back to 1982. Logic programming is mostly known outside the narrow circles through the programming language Prolog but there is much more to it than that, of course. Although we have not convinced the world that all programs should be written in Prolog or some its derivatives, logic programming has had a large influence in many areas such as databases, language processing, constraint solving and artificial intelligence.[1]

In general, logic has had an enormous influence on the development and use of computers—not only the simple analogy between the computer's most basic operations and boolean algebra, but because powerful theoretical frameworks are needed in order to meet the demand for more complex knowledge representation and manipulation, and for verification of critical software, etc. The other way round, the present day world with all its computers, advanced applications and needs for even more advanced applications have boosted new life to mathematical logic. There are plenty of thought-provoking aspects of computer applications that provide work for logicians.

Let me take one example. One of the workshops that I hope to have time to participate in is on Paraconsistent Computational Logic. Paraconsistency is an important issue if you consider databases, as I will argue: Databases are really the only kind of widespread software that provides a semantics that is associated with logical concepts. If all databases were consistent, first-order logic would be sufficient. When a database contains a single little piece of information that provokes logical inconsis-

[1] Artificial intelligence understood here not as attempts to simulate human intelligent behaviour but the more humble and pragmatic version of 'using present day computers to solve problems currently associated with human intelligence and on the border of what we usually expect computers capable of doing'. With this definition, Blaise Pascal's calculator was the artificial intelligence of his time.

tency, the theory says that the database can produce any kind of nonsense answers. However, any but toy databases have inconsistencies in them, but they still produce useful answers. Whattawedo as database logicians? Either concentrate on idealized small scale things or we start considering logics that give a less despairing handling of inconsistencies, i.e., paraconsistent logics. Practical issues that we can address if we can work with inconsistencies are to investigate to what extent we can trust answers from a database, how we can detect and measure inconsistency and how we can repair it. If you want to know more about this topic, follow the links from `http://floc02.diku.dk`.

Another workshop that interests me is on Natural Language Understanding and Logic Programming; my own personal work in this area concerns logical grammars based on constraint logic programming and combining it with abduction (a la Pierce). Roughly this means to make qualified guesses of which (to me initially unknown) real world facts that made the speaker deductively produce his or her utterance, thus providing me with an interpretation of what was said.

Here I have only mentioned a few examples of topics treated at FLoC'02—those things that I have some relations to. The field is very diverse and it is impossible for a single person to have a detailed knowledge and insight in all that goes on. This is the list of the conference titles at FLoC:

- Formal Methods Europe
  (July 22-24)
- Conference on Rewriting Techniques and Applications
  (July 22-24)
- IEEE Symposium on Logic in Computer Science
  (July 22-25)
- Conference on Automated Deduction
  (July 27-30)
- Conference on Computer-Aided Verification
  (July 27-31)
- International Conference on Logic Programming
  (July 29 - August 1)

- Automated Reasoning with Analytic Tableaux and Related Methods
  (July 30 - August 1)

The list of workshops which is too long to bring here can be found at FLoC's website. Full program and registration information for will also be available here.

6 DIMENSIONS IN EPISTEMIC LOGIC

▶ CONFERENCE—Roskilde University, May 2–4, 2002

Conference Aim

In the early 1960s Hintikka published his seminal work on the logic of knowledge and belief. Since then, epistemic and doxastic logics have grown into a mature discipline with many important applications in philosophy, computer science, game theory, economics and linguistics to mention a few areas of application. In philosophy, epistemic logic has contributed to the general epistemological understanding of propositional knowledge. Artificial intelligence and knowledge representation studies have profited from the tools, methods and systems of epistemic logic. Epistemic logic has also proved quite important in game theory and economics in terms of modelling for instance non-cooperative games of perfect information.

But epistemic logic suffers from a weakness, which it shares with many modal logics. Already in 1970 Dana Scott noted:

> Here is what I consider one of the biggest mistakes of all in modal logic: concentration on a system with just one modal operator. The only way to have any philosophically significant results in deontic logic or epistemic logic is to combine these operators with: Tense operators (otherwise how can you formulate principles of change?); the logical operators (otherwise how can you compare

the relative with the absolute?); the operators like historical or physical necessity (otherwise how can you relate the agent to his environment?); and so on and so on.

Even though this substantial criticism towards the way in which philosophical logic proceeds to a great extent still holds true, new results in epistemic logic suggest otherwise and begin to realize what Scott long has wished for. Recently logicians have begun to develop multi-modal systems in which both epistemic, alethic and temporal operators can be defined in unified formal frameworks. These multi-modal systems have then been used to study new features of our cognition like how knowledge may evolve over time, knowledge in linear versus branching time, the dynamics of knowledge databases, alethic modality and epistemic capabilities etc.

Dimensions in Epistemic Logic have to aims:

(i) To track the history and development of epistemic logic from Hintikka's first formulations to its contemporary forms and consider some of the many applications in philosophy, computer science, economics etc.

(ii) To describe and discuss the developments of epistemic logic in multi-modal systems.

All lectures will be of such a nature that they can be followed by students and scholars of philosophy, computer science, linguistics etc. without deep professional training in epistemic logic but provided with general knowledge of foundational issues.

Joseph Halpern

Computer Science Department, Cornell University, USA

▶ Substantive Rationality and Backward Induction

Some of the major puzzles in game theory today involve the notion of rationality. Assuming that all players are rational, and know that they are all rational, and know that they know, etc., results in strategies that seem highly irrational. At the last TARK (Theoretical Aspects of Rationality and Knowledge) conference, there was a 2.5 hour round table, involving some leading game theorists and philosophers, on 'Common knowledge of rationality and the backward induction solution for games of perfect information'. During the discussion Robert Aumann stated the following theorem:

> *Common knowledge of substantive rationality implies the backward induction solution in games of perfect information.*

Robert Stalnaker then stated the following theorem:

> *Common knowledge of substantive rationality does not imply the backward induction solution in games of perfect information.*

In this talk I will carefully explain all the relevant notions (games of perfect information, knowledge and common knowledge, strategies, rationality, and substantive rationality) and explain why, although both Aumann and Stalnaker were apparently using the same definitions, they were able to (correctly) prove such different results. The key turns out to lie in getting a good model of counterfactual reasoning in games. I will in fact provide a formal model that allows us to prove both results and to understand the technical differences between them. The model has the added advantage of giving us a deeper insight into what conclusions we can draw from rationality and common knowledge of rationality. No prior knowledge will be assumed.

Jaakko Hintikka

Department of Philosophy, Boston University, USA

▶ Independence—the most important dimension of epistemic logic

The first generation epistemic logic promised several important insights, among them an analysis of different kinds of knowledge (knowledge of facts vs. knowledge of objects), a logic of questions and answers, the role of mathematical and other conceptual knowledge in empirical inquiry and a strategic parallelism of deduction and interrogation that shows the role of logic in epistemology. On a closer examination, it turned out that these insights applied only to the simplest kinds of knowledge. By using game-theoretical semantics and the notion of independence, a second generation epistemic logic can be reached that enables us to define all the crucial concepts for all types of knowledge and for all kinds of questions. This improved epistemic logic thus fulfills the promises that the first generation epistemic logic suggested but did not fully keep.

Wiebe van der Hoek

Department of Computer Science, University of Utrecht, The Netherlands

▶ On Epistemic Action Languages

Information pervades so many aspects of our daily lives that the age we live in has been dubbed 'the information age'. Epistemic logic studies information with an emphasis on 'information about information'. The concept of common knowledge is an excellent example of this how 'information about information' works. Recently a new field of epistemic logic has developed: 'dynamic epistemic logic'. It studies how information changes, especially information about information. Several systems have been put forward. In our paper we explore a new approach to dynamic epistemic logic.

In the late nineties, several proposals have been made to model actions with epistemic aspects. In most of them, actions are seen as multi-agent Kripke models, where instead of a

valuation of propositional variables there is a precondition function. This provides a good idea for the semantics of a dynamic epistemic logic, because it yields an interesting class of actions. The question is which syntax is appropriate to describe these actions. In our paper we discuss the approaches that have been put forward. We also introduce a new language to reason about these actions, to overcome some of the problems of the other proposals.

Joint work with Barteld Kooi and Hans van Ditmarsch.

Wolfgang Lenzen

Department of Philosophy, University of Osnabrück, Germany

▶ Knowledge, Belief, and Subjective Probability—Outlines of a Unified System of Epistemic/Doxastic Logic

My paper is divided in three parts. In the first (and main) part I will give a survey of what has been achieved in the early days of epistemic logic. In the second part I want to comment upon what had then been considered as the open problems for further research in epistemic logic. In the third part I will briefly touch upon what I now consider as the most important desiderata of future research in epistemic logic.

▶ Extended Abstract

1 The Early Days

I begin by summarizing the main results and achievements of Jaakko Hintikka's work in epistemic logic. In his pioneering book *Knowledge and Belief* (1962) Hintikka laid the foundations of what is now known as epistemic logic by formalizing certain epistemological principles for the basic relations 'a knows that p' and 'a believes that p' and by developing a semantic framework for the logical interpretation of these concepts. Moreover, in various papers collected in *Models for Modalities* (1969) and *The Intentions of Intentionality* and other new models for modalities (1975), Hintikka investigated both the parallels and the differences between epistemic modalities on the one hand and certain other modal operators on the other hand. In

particular, he helped to clarify the intricate problems involved in quantifying in modal contexts.

My own study of *Recent Work in Epistemic Logic* (1978) was intended as a critical survey of the (real and apparent) problems in epistemic logic as they were discussed in the literature from 1960 onwards. I argued in particular that many problems disappear if sufficient care is taken to distinguish between *semantical* issues pertaining to the *truth-conditions* for epistemic attitudes and *pragmatic* issues pertaining to the conditions for the rational *utterability* of epistemic propositions or sentences. The aim of my 1980 book *Glauben, Wissen und Wahrscheinlichkeit* was to bring up doxastic/epistemic logic to the standards of other branches of philosophical logic. On the one hand I attempted to show that the logic of the knowledge-operator is isomorphic to the system S4.2 of alethic modal logic. On the other hand I argued that studies in the logic of belief should best be based upon the theory of subjective probability. This meant in particular that one has to distinguish between at least two different notions of belief, so-called weak and strong belief, where someone weakly believes that p iff he considers p as likely or *probable*, say, in the sense of $Prob(a,p) > \frac{1}{2}$, while a strongly believes that p iff a is entirely *convinced* that p, i.e. $Prob(a,p) = 1$. This approach resulted in a combined system of doxastic/epistemic logic where the three operators $K(a,p)$ 'a knows that p', $B(a,p)$ 'a believes that p', and $C(a,p)$ 'a is convinced that p' are interconnected by various laws, e.g.:

$$C(a,p) \leftrightarrow B(a,K(a,p)).$$

2 Revisions and Refinements

The most controversial issues discussed in those old days were

- the problem of so-called 'logical omniscience' (and, similarly, logical omni-belief);
- the problems concerning the legitimacy of 'quantifying in' epistemic propositions, including the important distinction between epistemic attitudes *de dicto* and *de re*;

- the issue, raised in particular by H.-N. Castaneda, of the proper analysis of epistemic attitudes *de se*, i.e. opinions of the subject a concerning properties of *his own*.

In my opinion the last two problems may be regarded as solved, while the first problem heavily depends on the methodological role which one wants to assign to epistemic logic in general. I will briefly discuss the main options and explain why a recourse to non-standard logics such as, e.g., relevance logic, or to non-standard semantics, such as an impossible worlds-semantics, is not very promising.

3 Future Developments

I didn't follow up the literature on epistemic logic in the 80ies and 90ies very intensively, so my comment on recent developments is bound to remain somewhat superficial. Judging basically from what I learned at the series of biennial conferences on *Theoretical Aspects of Reasoning about Knowledge* from 1986 onwards, there was a trend towards

- an integration of epistemic modalities into systems of *multi-dimensional* modal logic;
- a generalization of the 'solipsistic' systems describing, e.g., the structure of the belief-system of *one* person a, to systems considering also a's belief about *others*' beliefs (including systems dealing with the common belief of a group of people, $G = \{a, b, c, \}$).

I hope that the present conference will open new perspectives on these tasks as well as on what I personally find the most important goal of future research in epistemic logic, *viz.*:

- building a bridge between the 'static' systems of ordinary epistemic logic on the one hand and 'dynamic' systems like that of Gärdenfors (1988) designed to model ('rational') *belief-change* on the other hand.

Hans Rott

Department of Philosophy, University of Regensburg, Germany

▶ Interpreting Belief Revision

What is 'economic' in the theory of belief change?

We must justify the construction of ever new logical systems – including systems of epistemic logic and systems of belief dynamics – by giving as substantial motivation as possible for the systems we develop. In this talk, I ask to what extent belief revision may be regarded as a branch of cognitive economics. For a long time, beginning at the latest with the publication of Gärdenfors's seminal *Knowledge in Flux* in 1988, belief dynamics has been said to be driven primarily by a concern for 'informational economy', a maxim also known as 'conservatism'. We take a close look at the role that informational economy has actually played in the development of belief change theories in the last two decades, and ask whether it should take a (more) important role in their motivation. We discuss the idea of conservatism both with respect to beliefs and with respect to richer belief states (identified with belief-revision guiding structures). Our finding is negative for both respects, and it is negative both descriptively and normatively. This view of cognitive economics is then contrasted with an alternative one taking belief change to be a problem of rational choice, more precisely, of choice based on complete and transitive preferences. It turns out that under this interpretation, belief revision theory is indeed amenable to an essentially economic interpretation, but that it inherits criticism that has been levelled against the classical theory of choice in wider contexts.

Krister Segerberg

Department of Philosophy, Uppsala University, Sweden

▶ Two Attempts at Deconstructing Epistemic Logic

Traditional epistemic and doxastic logic, introduced by Jaakko Hintikka, is the logic of knowledge and belief of agents who are perfect in these sense that they know all logical consequences of what they know, and believe all logical consequences of what they believe. It is often suggested that Hintikka's logic is in

some sense a limiting case of logics of agents whose powers of reasoning are more limited. But to work out that suggestion in an interesting way is not easy. In particular, it is a challenge to try to come up with interesting logics of knowledge and belief of agents whose reasoning powers are much less than those of Hintikka's agents. In this talk I shall present some (not entirely successful) efforts at trying to meet that challenge.

John F. Sowa

▶ Laws, Facts, and Contexts: Foundations for Multimodal Reasoning

Leibniz's intuition that necessity corresponds to truth in all possible worlds enabled Kripke to define a rigorous model theory for several axiomatizations of modal logic. Unfortunately, Kripke's model structures lead to a combinatorial explosion when they are extended to all the varieties of modality and intentionality that people routinely use in ordinary language. As an alternative, any semantics based on possible worlds can be replaced by a simpler and more easily generalizable approach based on Dunn's semantics of laws and facts and a theory of contexts based on the ideas of Peirce and McCarthy. To demonstrate consistency, this paper defines a family of nested graph models, which can be specialized to a wide variety of model structures, including Kripke's models, situation semantics, temporal models, and many variations of them. An important advantage of nested graph models is the option of partitioning the reasoning tasks into separate metalevel stages, each of which can be axiomatized in classical first-order logic. At each stage, all inferences can be carried out with well-understood theorem provers for FOL or some subset of FOL. To prove that nothing more than FOL is required, Section 6 of this paper shows how any nested graph model with a finite nesting depth can be flattened to a conventional Tarski-style model. For most purposes, however, the nested models are computationally more tractable and intuitively more understandable.

▶ Extended Abstract

(i) Replacing Possible Worlds with Contexts

1 Replacing Possible Worlds with Contexts

Possible worlds have been the most popular semantic foundation for modal logic since Kripke (1963) adopted them for his version of model structures. David Lewis (1986), for example, argued that 'We ought to believe in other possible worlds and individuals because systematic philosophy goes more smoothly in many ways if we do.' Yet computer implementations of modal reasoning replace possible worlds with 'ersatz-worlds' consisting of collections of propositions that more closely resemble Hintikka's (1963) model sets. By dividing the model sets into necessary laws and contingent facts, Dunn (1973) defined a compatible refinement of Kripke's semantics that eliminated the need for a 'realist' view of possible worlds. Instead of assuming Kripke's accessibility relation as an unexplained primitive, Dunn derived it from the selection of laws and facts. Since Dunn's semantics is logically equivalent to Kripke's for conventional modalities, most logicians ignored it in favor of Kripke's. For multimodal reasoning, however, Dunn's approach enormously simplifies the reasoning processes. In effect, it enables a clean separation of the metalevel reasoning about the choice of laws and facts from the object-level reasoning in pure first-order logic.

To take advantage of Dunn's semantics, the metalevel reasoning should be performed in a separate context from the object-level reasoning. This separation requires a formal theory of contexts that can distinguish different metalevels. But as Rich Thomason (2001) observed, the theory of context is important and problematic—problematic because the intuitions

are confused, because disparate disciplines are involved, and because the chronic problem in cognitive science of how to arrive at a productive relation between formalizations and applications applies with particular force to this area. The version of contexts adopted for this paper is based on a representation that Peirce introduced for his existential graphs and Sowa (1984) adopted as a foundation for his version of conceptual graphs (CGs). That approach is further elaborated along the lines suggested by McCarthy (1993) and developed by Sowa (1995, 2000).

Sections (ii), (iii), and (iv) of this paper summarize Dunn's semantics of laws and facts, a theory of contexts based on the work of Peirce and McCarthy, and Tarski's hierarchy of metalevels. Then Section (v) introduces nested graph models (NGMs) as a general formalism for a family of models that can be specialized for various theories of modality and intentionality. Section (vi) shows how any NGM with a finite depth of nesting can be flattened to a Tarski-style model consisting of nothing but a set I of individuals and a set R of relations over I. Although the process of flattening shows that the modalities can be represented in first-order logic, the flattening comes at the expense of adding extra arguments to each relation to indicate every context in which it is nested. Finally, Section (vii) shows how reasoning about multiple modalities can be performed in and about the contexts and the propositions nested within them.

Robert Stalnaker

Department of Philosophy, MIT, USA

▶ Knowledge and Belief, Rational Action and Interaction: Epistemic models for games

Interactive epistemology — the application of epistemic logic and semantics to game theory — throws light both on some patterns of strategic reasoning (such as backward induction reasoning) and also on some general questions about the nature of knowledge and the relation between knowledge and belief. Specifically, the effect of the assumption that the players in a game have common knowledge that they all are rational depends

on exactly what is assumed about knowledge. In this talk, I will explore some different assumptions about knowledge, and trace some of their consequences in a game theoretic setting. I will suggest that the application of epistemic logic in a theory of practical strategic reasoning helps to connect formal questions about the logic of knowledge with some classical problems of substantive epistemology.

Moshe Y. Vardi

Department of Computer Science, Rice University, USA

▶ Common Knowledge Revisited

The notion of common knowledge, where everyone knows, everyone knows that everyone knows, etc., has proven to be fundamental in various disciplines, including Artificial Intelligence, Economics, Game Theory, Psychology, and Distributed Computer Systems. Common knowledge seems to present us with a paradox. On one hand, common knowledge is necessary for agreements and for coordination. On the other hand, common knowledge is unattainable in the real world. In this expository talk we will describe the paradox and various approaches to its resolution.

This talk represents joint work with R. Fagin, J.Y. Halpern, and Y. Moses.

Ryszard Wojcicki

Department of Philosophy and Sociology, Polish Academy of Sciences, Poland

▶ Referential Semantics

The truth value that a formula acquires may be different on different occasions (e.g. in different contexts or in different possible worlds). Call any such an occasion a reference point. If there are only two truths values (truth and falsity), the class of logics that are strongly complete (i.e. entailment relation coincides with that of logical derivability) with respect to referential semantics is the same as that of those logics that have the

following syntactical property: If two formulas A and B are logically equivalent, so are any two formulas of the form $C(...A...)$, $C(...B...)$. After presenting the technical part in a rather short way, I shall focus on epistemic aspects of the discussed ideas.

Conference Chair

- Vincent F. Hendricks, vincent@ruc.dk
- Stig Andur Pedersen, sap@ruc.dk

Location and Time Tables

All presentations will take place at Roskilde University (RUC) in the auditorium in building 46. On the back of this booklet, please find a map of the University. To reach RUC at 8:45 take a train from the Central Station of Copenhagen at 8:10 to Roskilde and get off at Trekroner Station. There will be signs from the station. To reach RUC from Roskilde at 8:52 take a bus from Roskilde Station at 8:41. (An appropriate train for Saturday leaves Copenhagen Central Station at 8:34).

	Thursday, May 2
9:00 - 9:20	Registration
9:20 - 9:30	Opening
9:30 - 10:30	J. Hintikka
10:30 - 11:00	Discussion
11:00 - 11:30	Coffee Break
11:30 - 12:30	J. Halpern
12:30 - 13:00	Discussion
13:00 - 14:00	Lunch
14:00 - 15:00	W. v. d. Hoek
15:00 - 15:30	Discussion
15:30 - 16:00	Coffee
16:00 - 17:00	Sight-seeing
17:00 - 18:00	tour

Conference Chair: Vincent F. Hendricks

	Friday, May 3
9:00 - 9:20	
9:20 - 9:30	
9:30 - 10:30	W. Lenzen
10:30 - 11:00	Discussion
11:00 - 11:30	Coffee Break
11:30 - 12:30	K. Segerberg
12:30 - 13:00	Discussion
13:00 - 14:00	Lunch
14:00 - 15:00	H. Root
15:00 - 15:30	Discussion
15:30 - 16:00	Coffee
16:00 - 17:00	R. Wojcicki
17:00 - 17:30	Discussion

Conference Chair: Stig Andur Pedersen

	Saturday, May 4
9:00 - 9:20	
9:20 - 9:30	
9:30 - 10:30	R. Stalnaker
10:30 - 11:00	Discussion
11:00 - 11:30	Coffee Break
11:30 - 12:30	M. Vardi
12:30 - 13:00	Discussion
13:00 - 14:00	Lunch
14:00 - 15:00	J. F. Sowa
15:00 - 15:30	Discussion
15:30 - 16:30	Panel Discussion
16:30 - 16:40	Closing

Conference Chair: Vincent F. Hendricks & Stig Andur Pedersen

Lunch and Dinner Arrangements

On Thursday, Friday and Saturday participants may choose to order lunch through the conference organization; payments for these arrangements are due during final conference registration on Thursday May 2nd. Each lunch costs 50,00 Dkr and includes besides Danish 'smørrebrød' (open sandwiches) one beverage. Registered participants who do not wish to order lunch may either bring their own or buy lunch from the campus canteen. Note, however, that the canteen is closed on Saturday. A conference dinner is scheduled for Friday, May 2nd at 19:00, Restaurant Gråbrødre Torv in the center of Copenhagen. The dinner costs 350,00 Dkr and includes a three course meal and wine. The student price is reduced to 175,00 Dkr. Only a limited number of seats are available. Participants interested in lunch orders and/or conference dinner participation should notify ΦLOG-secretary Pelle Guldborg Hansen (see address below) no later than Friday, April 26th. Only *cash* payments are accepted and no later than upon final conference registration

during Thursday, May 2nd.

Sight-seeing Tour

A sight-seeing tour of Copenhagen is planned for Thursday, May 2nd. The tour will cost 50,00 Dkr which includes transportation, guides and passes. Signing up for the tour is, as for the lunch and dinner arrangements, no later then Friday, April 26th, and payment is due no later than upon final registration.

Registration

Registration is free. Please write the ΦLOG secretary Pelle Guldborg Hansen:

Pelle Guldborg Hansen
Department of Philosophy and Science Studies
Roskilde University, PA6
P. O. Box 260
DK4000 Roskilde, Denmark
Phone (+45) 4674 2343 Fax (+45) 4674 3012
Email pgh@ruc.dk

Please be sure to include your name, institution, country and zip-code and your email address.

If email is used include EPISTEMIC REGISTRATION in the subject entry. All questions pertaining to registration and accommodations should be directed to Pelle Guldborg Hansen. No individual notification upon registration will be forwarded to individual participants. An updated list of participants may be found by consulting the ΦLOG homepage at

`http://www.philog.ruc.dk`

7 ΦLOG OG SHF

► IN DANISH

ΦLOG er resultatet af en ansøgning til Statens Humanistiske Forskningsråds netværksbevillinger. Ansøgningen blev indsendt i oktober 2001 og det positive svar forelå i december 2001. Nedenfor er bragt et uddrag af den oprindelige ansøgning udfærdiget af Vincent F. Hendricks og Stig Andur Pedersen med støtte fra ΦLOG's koordinationsgruppe.

Baggrund

Filosofisk logik spiller i dag en væsentlig rolle ikke blot i filosofi, men tillige i datalogi, sprogvidenskab, argumentationsteori, jura og psykologi for blot at nævne nogle vigtige områder. Den filosofiske logik dækker mange forskelligartede logiske studier og aktiviteter i humaniora og dens grænseflade til naturvidenskab:

(i) Modal logik (eller aletisk logik).
Studiet af måderne, eller modaliteterne, hvormed forskellige udsagnstyper kan være sande eller falske.

(ii) Temporal logik.
Det formelle studie af tiden, dens struktur og egenskaber.

(iii) Epistemisk, doxastisk logik og vidensrepræsentation.
Studiet af vor viden, overbevisning og erkendelse, deres udvikling, struktur, styrke og gyldighed.

(iv) Deontisk logik og etisk/juridisk ræsonnering.
Studiet af den logiske struktur af moralske påbud og vurderinger samt studiet af etiske/juridiske ræsonnæringsformer.

(v) Logikprogrammering.
Udvikling af systemer til modellering af ræsonnæringsformer, udvikling af databaseapplikationer etc.

Denne liste er ikke udtømmende, men det er klart, at den filosofiske logik har et stort virkeområde og indgår i et tværfagligt forskningsfelt i spørgsmål om at forstå viden, vidensproduktion, argumentation, kognitive og lingvistiske modeller etc.

Danmark har fagligt set meget højt kvalificerede og internationalt anerkendte forskere indenfor filosofisk logik, der alle har indvilliget i at deltage i netværket:

(i) Modal logik:
Per Hasle (SDU), Peter Øhrstrøm (AUC), Torben Bräuner (RUC), Lars Bo Gundersen (AAU), Vincent F. Hendricks (RUC), Stig Andur Pedersen (RUC).

(ii) Temporal logik:
Per Hasle, Peter Øhrstrøm, Torben Bräuner, Lars Bo Gundersen, Vincent F. Hendricks, Stig Andur Pedersen.

(iii) Epistemisk og doxastisk logik:
Jan Riis Flor (KU), Lars Bo Gundersen, Torben Bräuner, Vincent F. Hendricks, Stig Andur Pedersen, Cynthia Grund (OU-SDU)

(iv) Logikprogrammering:
Henning Christiansen (RUC), Torben Bräuner, Per Hasle, Peter Øhrstrøm.

Begrundelse for netværket

Der findes flere gode grunde til at oprette et netværk, som koordinerer og integrerer de forskellige logiske og filosofiske forskn-

ingsaktiviteter i landet:

► Systematisering og koordinering af forskningsmiljøerne i Danmark

Som det fremgår af ovenstående, har Danmark en række forskere, der arbejder indenfor filosofisk logik. Imidlertid arbejder disse forskere, selvom de fleste er bekendt med de øvrige forskeres arbejde, isoleret fra hverandre. Et netværk i filosofisk logik ville kunne åbne de forskellige forskeres arbejdsområder for hinanden, lade netværkets deltagere gensidigt inspirere og udveksle ikke blot forskningsresultater og idéer men lige såvel internationale kontakter. Et netværk i filosofisk logik ville således kunne danne rammen om en frugtbar interaktion og synergi mellem forskere, forskningsområder og den interna-tionale forskningsverden.

Der findes på nuværende tidspunkt et tilsvarende netværk indenfor matematikkens filosofi og historie, MATHNET, med støtte fra Statens Naturvidenskabelige Forsknings-råd. Siden dette netværks start i 1996 har MATHNET med stor succes formået at systematisere, koordinere og konsolidere den samlede danske forskning indenfor dette felt. Se ydermere

`http://www.mathnet.ruc.dk.`

Idéen med ΦLOG er herefter at oprette et netværk af tilsvarende karakter og forhåbentlig med tilsvarende succes.

► Behovet for en kvalificeret diskussion af den filosofiske logiks rolle

Den filosofiske logik har spillet en væsentlig rolle i en række forskellige discipliner: Eksempelvis er store dele af den indflydelsesrige filosof, Hilary Putnams, videnskabs- og erkendelsesteori inspireret og afledt af grundlæggende logiske resultater [Putnam 81]. Samspillet mellem førsteordens logikken og behaviorismen har været væsentlige faktorer i W. v. O. Quines erkendelsesteori [Quine 75]. Studiet og forståelsen af induktive problemer og deres løselighed har også udviklet sig markant ved hjælp af logiske metoder hentet fra den formelle indlæringsteori og dens tilhørende beregnelighedsstudier [Kelly 96]. Formel indlæringsteori spiller ydermere en afgørende rolle indenfor lingvistikken i modelleringen af sprogindlæring [Osherson, Stob, Weinstein 86]. I etik og metaetik har Von Wright [Von Wright 51]

bidraget til forståelsen af moralske påbud/vurderinger og deres logiske struktur og semantik ved hjælp af deontisk logik. Deontisk logik har senere sammen med den såkaldte non-monotone logik givet anledning til frugtbare modelleringer af juridisk ræsonnering.

I datalogien og i særdeleshed indenfor logikprogrammeringen har den filosofiske logik ligeså vel spillet en markant rolle. Vidensrepræsentation og vidensudtræk i implementeringsmæssigt øjemed har set en frugtbar udvikling via anvendelse af formalismer hentet fra både den epistemiske og doxastiske, temporale og modale logik samt teorien om belief revision [Gärdenfors 88]. Dertil kommer, at databaseapplikationer og programmeringssprog har fundet stor anvendelse for resultater fra den del af filosofisk logik, der vedrører kontrafaktiske konditionaler, parakonsistente logikker og vedligeholdelse af konsistens i databaser.

Siden netværket, givet dets deltagere, har faglig bredde som spænder fra filosofi og sprogvidenskab til datalogi, er det i stand til at etablere en kvalificeret diskussion, udveksling og udvikling af den filosofiske logik.

► Behovet for at få den filosofiske logik ud til en bredere offentlighed

Logik har i mange år ikke været en særlig dyrket disciplin i Danmark. Det kommer sig måske af, at mange har haft svært ved at se logikkens anvendelighed i mere generelle eksempelvis filosofiske, sprogvidenskabelige, argumentationsteoretiske, datalogiske og juridiske spørgsmål. På den anden side, så er den filosofiske logik, til forskel fra den rene eller matematiske logik, netop i udgangspunktet karakteriseret derved, at den som genstandsområde har et bredere perspektiv. Et netværk i hvilket forskere med forskellige interesser og udgangspunkter, der dog alligevel arbejder indenfor området, ville således kunne formidle resultater på tværs af discipliner og faggrænser og således skærpe den generelle interesse og opmærksomhed for filosofisk logik i Danmark.

Nyere tendenser i filosofisk logik

Filosofisk logik er på sin vis ikke en ny disciplin forstået på den måde, at man allerede i antikken var bekendt med især modale og temporale udsagnstyper, kontekster og argumenter. Op igennem historien finder man også spredte studier af filosofisk logisk karakter. Systematiske og formelle studier af filosofisk logik er imidlertid noget, der for alvor tog fart i sidste århundrede. Grundlæggende syntaktiske repræsentationer, semantiske modeller og logiske systemer blev udviklet indenfor både modal, temporal, epistemisk, doxastisk og deontisk logik [Hughes & Cresswell 96], [Chellas 80], [Prior 67], [Hintikka 62], [Von Wright 51]. Efter disse grundlæggende studier har såvel datalogien, lingvistikken og psykologien på værdifuld vis såvel bidraget til som draget fordel af disse fundamentale systemer i særdeleshed i forbindelse med logikprogrammering indenfor databaseapplikationer, computerlingvistisk modellering og kognitionsteoretiske modeller, vidensrepræsentation og produktion.

Allerede i 1970 påpegede den indflydelsesrige logiker Dana Scott imidlertid følgende svaghed i den filosofiske logik og den måde, hvorpå den i dag studeres og anvendes:

> Here is what I consider one of the biggest mistakes of all in modal logic: concentration on a system with just one modal operator. The only way to have any philosophically significant results in deontic logic or epistemic logic is to combine these operators with: Tense operators (otherwise how can you formulate principles of change?); the logical operators (otherwise how can you compare the relative with the absolute?); the operators like historical or physical necessity (otherwise how can you relate the agent to his environment?); and so on and so on. [Scott 70], p. 143.

Scotts centrale pointe er således den, at selvom de forskelligartede logikker både er ganske velartikulerede, velforståede og nyder betragtelig udbredelse og anvendelse i meget forskellige discipliner, så er de velartikulerede og velforståede isoleret set og hver for sig. Som det dog fremgår af citatet, er der behov

for nye formelle systemer, der kan håndtere flere logikker på én og samme gang i et og samme system.

I den internationale forskningsverden har man nu erkendt Scotts indsigt, men kun meget få er begyndt at komme med bud på, hvorledes et sådan eller sådanne systemer bør tage sig ud. [Fagin et al. 95] har blandet epistemisk logik og temporal logik for at studere videns udvikling i tid i multi-agent systemer med opmærksomhed for datalogisk implementering. Blandt andet [Zanardo 96] har studeret forholdet mellem modallogik og temporal logik.

Der er imidlertid danske forskere, der er helt på forkant med denne nye tendens. Både på AUC, SDU og RUC har forskere knyttet til netværket arbejdet indgående med forholdet mellem modal- og temporal logik. Forskere på RUC har tillige udviklet et system i hvilket modallogiske, epistemiske og temporale logikker kan defineres på én og samme tid. Et netværk i filosofisk logik, der blandt andet som ambition har at være medvirkende årsag til udviklingen af sådanne generelle og forenende formelle systemer, vil for alvor placere Danmark på det internationale landkort i filosofisk logik.

En anden tendens indenfor filosofisk logik er knyttet til datalogien, der viser sig i to dimensioner, som i de senere år udviser et større og større overlap: (1) studier omkring formalisering og implementation af filosofiske og sproglige begreber [Peirce 58] og [Stalnaker 91] baseret på logik- og constraintprogramering. (2) nye udviklinger indenfor grundlæggende teknologiske områder som programmeringsssprog, databaser, natursprogs-analyse og vidnesudtræk. Emner som abduktion, induktion, kontrafaktisk ræsonnering, etablering og vedligeholdelse af konsistens dukker naturligt op i dette krydsfelt. En datalog på RUC, der også er knyttet til netværket har udviklet metoder, der kan få en og samme bevismaskine til at udføre deduktion, abduktion og (i mindre omfang) induktion som en fælles integreret proces. Nært relateret hertil er også studiet af og implementation af forskellige former for negation i data- og vidensbaser. Nyt arbejde går ud på at anvende constraint-logik til robust bottom-up-analyse af naturligt sprog, som på naturlig måde benytter abduktion, integritetsbegrænsninger, og muligvis også aspekter af lineær logik til håndtering af anaforer.

Formålet med netværket

Netværkets vigtigste aktivitet vil være at iværksætte og koordinere ny forskning inden for filosofisk logik og dens anvendelser. Det primære formål vil være at få en systematisk forståelse af de nye tendenser i filosofien, lingvistikken og datalogien og deres formelle repræsentation og egenskaber. Dertil kommer at bidrage til denne forskning ved udførelse af konkrete filosofiske og logiske studier. Flere af de forskere, der er tilknyttet netværket, er allerede engageret i arbejdet og besidder betydelige kontakter i centre og universiteter i USA og Europa, som der kan trækkes på i forbindelse med forskellige aspekter af forskningsprogrammet.

Foruden at intensivere og iværksætte forskning vil det også være en vigtig opgave for netværket at arrangere kurser og møder, som både kan støtte forskningen og bidrage til at formidle resultater ud til en bredere kreds. Specielt vil det være vigtigt at formidle resultater til undervisere og forskere i filosofi, datalogi, lingvistik, psykologi på såvel de højere læreranstalter og i private forskningsvirksomheder, sektorforskningsinstitutioner etc. Flere møder vil blive planlagt på en sådan måde, at de ikke kun henvender sig til specialister, men også til en skare af logik- og filosofi- interesserede og en bredere offentlighed.