Jaap Kamps

THE UBIQUITY OF BACKGROUND KNOWLEDGE

ABSTRACT. Scientific discourse leaves implicit a vast amount of knowledge, assumes that this background knowledge is taken into account—even taken for granted—and treated as undisputed. In particular, the terminology in the empirical sciences is treated as antecedently understood. The background knowledge surrounding a theory is usually assumed to be true or approximately true. This is in sharp contrast with logic, which explicitly ignores underlying presuppositions and assumes uninterpreted languages. We discuss the problems that background knowledge may cause for the formalization of scientific theories. In particular, we will show how some of these problems can be addressed in the context of the computational representation of scientific theories.

1 *Introduction*

Background knowledge is ubiquitous in all forms of meaningful human communication. People engaged in fruitful discussion rely on a vast amount of shared background knowledge. How can we communicate if, for example, we do not have a shared understanding of the meaning of the words we utter? or make the same underlying assumptions? I vividly recall a discussion with Professor Kuipers on our common interests in artificial intelligence and philosophy of science. After much agreement, we suddenly reached an awkward difference of opinion that left me puzzled for some time. Then it turned out to be the case that Professor Kuipers was talking about the beneficial effects philosophy of science can have on artificial intelligence, and I was talking about the beneficial effects artificial intelligence can have on philosophy of science. Since these two positions are by no means incompatible, our difference of opinion was immediately resolved. This anecdote illustrates how a minor difference in the implicit presuppositions can give rise to confusion and even apparent disagreement, and moreover, how this may be resolved after the background assumptions have been made explicit. Background knowledge does not only occur in free forms of conversation, but also in more regulated discourse we make all sorts of presuppositions. This is even true for the way in which we report our findings and theories in the scientific literature. That is, even in cases where the clarity and unambiguity is of principal importance, authors routinely presuppose a variety of background knowledge, for example, by the terminology that they use.

The notion of 'background knowledge' is traditionally used to denote the vast amount of knowledge we take for granted when discussing a problem; this knowledge is treated as undisputed, if only for the time being and for the problem at hand (Popper 1963, p.238). If some parts of the background knowledge are called into question, they do no longer belong to the background knowledge. As Kuipers (2001, p.6) puts it: "It should also be stressed that, at least as a rule, observation and hence observation terms are, and remain, laden by theoretical presuppositions which are considered to belong to the so-called unproblematic background knowledge." The background knowledge is "unproblematic" in the sense that we (have to) assume that it is true or approximately true (Kuipers 2001, p.48/p.51). A well-known consequence of the background knowledge surrounding a theory is the Duhem-Quine thesis, i.e., the observation that we can make a theory immune for falsification by making modifications in the background knowledge (Kuipers 2001, p.225/p.244). This paper discusses the problems that background knowledge may cause for the formalization of scientific theories. In particular, we will address these problems in the context of computational representation of scientific theories.

Due to the fact that background knowledge is taken for granted, it will remain implicit in written expositions of a theory, and only the relevant knowledge is mentioned. The explicit treatment of underlying assumptions is one of the main reasons for the formalization of scientific theories (Suppes 1968). Of course, one may argue that the background assumptions that are left implicit are often relatively innocuous, and frequently the authors may safely assume that these implicit assumptions belong to the common knowledge of the readers. However, if our goal is to provide a version suitable for computational reasoning, this assumption is no longer valid. Computers are simply not endowed with this underlying background knowledge, and all relevant implicit assumptions need to be added explicitly. This gives rise to several problems. The first is a problem of acquisition: how to bring to light the knowledge that has been left implicit? The second is a problem of relevance: the amount of implicit background knowledge seems without an end, how to decide which part of it are relevant for the problem at hand?

The problem of background knowledge will occur in any situation where there is prior knowledge at stake, including all of the empirical sciences. Regardless of the used representation language, there will always be the question of whether one has faithfully represented the presuppositions of the domain. In order to explicate the role of background knowledge in the formalization of theories, we will situate our discussion in the context of the axiomatization in first-order logic of theories from the empirical sciences. Our experience in this area concerns the informal theories of fields like sociology rather than the mathematical theories of physics.[1]

[1] This is roughly based on some recent attempts to axiomatize informal sociological theory (Péli et al. 1994; Hannan 1998; Kamps and Pólos 1999). Although one may expect that the more rigorous and formal an exposition is, the more of the background assumptions have been added explicitly and that the more informal an exposition is, the greater the amount of background knowledge that is presupposed. As a consequence, one would expect that, relative to the explicitly discussed part,

2 *Background Knowledge and Interpreted Languages*

Suppose that we start out with a conventional exposition of a scientific theory, think of an article appearing in a scientific journal. A careful rational reconstruction of such a text will result in a list of statements representing the axioms of the theory, and a list of statements representing the claims or predictions of the theory. This rational reconstruction is by no means a trivial step, but we will ignore these complications and assume that, at least for some texts, it can be accomplished. As a next step, we would want to give a formal rendition of the selected statements, and thus construct an initial formal version of the theory. This initial formal theory, which we assume here to be in first-order logic, will have a number of axioms and a set of conjectures representing the statements that the theory claims to predict or explain. We can now try to find out which of these conjectures can be derived from the axioms. In particular, we can use the standard tarskian consequence relation by using standard rules of inference (see standard textbooks like Enderton 1972). As may come as no surprise, this will generally be a disappointing effort: in the (initial) formal theory many of the conjectures will not be derivable from the axioms. As is well-known, informal arguments do not straightforwardly extend to rigorous formal proofs. Admittedly, in some cases this might be due to infelicitous argumentation. Some of the informal conjectures may turn out to be false when subjected to greater scrutiny. However, more generally speaking, there are other reasons for a failure to derive some of the informal conjectures.

In particular, one may question whether the standard consequence relation is faithfully singling out the intended consequences of our theory. As (Tarski 1946, pp.121–122) put it:

> Our knowledge of the things denoted by the primitive terms ... is very comprehensive and is by no means exhausted by the adopted axioms. But this knowledge is, so to speak, our private concern which does not exert the least influence on the construction of our theory. ... We disregard, as is commonly put, the meaning of the primitive terms adopted by us, and direct our attention exclusively to the form of the axioms in which these terms occur.

Now consider the empirical science theory we are axiomatizing: it contains primitive terminology that has a specific meaning—it is 'antecedently understood.' The standard logical consequence relation does not take into account the underlying understanding of the terminology. In other words, by using a standard consequence relation we explicitly ignore all background knowledge and assume that there are no logical relations among the atomic sentences other than those explicitly stated in the axioms. This is in sharp contrast with the discussion of unproblematic and undisputed background knowledge, which assumes that the antecedent meaning

authors in social sciences would leave larger parts of their theories implicit than is the case in, for example, mathematical physics. However, this is only a difference in degree, and does not affect the main points of our arguments.

of the used terminology is taken into account—even taken for granted. The unavoidable conclusion is that the failure to derive some of the informal conjectures can be attributed to the false assumption that we are dealing with an uninterpreted language (by using a tarskian consequence relation). In particular, some of the informal conjectures might materialize into formally proven theorems, were we to use a consequence relation that takes the underlying interpretation of terminology into account.

This has some far-reaching consequences. It is simply incorrect to regard our initial formal theory as an uninterpreted first-order language, but it should be regarded as an interpreted first-order logic in which the vocabulary has a specific, fixed interpretation. A direct result of using an interpreted language is that we cannot use the standard consequence relation. This requires a non-standard consequence relation that takes into account the interpretation of the terminology in the vocabulary of the theory. That is, for every set of interpreted vocabulary we need a special consequence relation that takes the antecedent meaning of the terminology into account.[2] In order to decide whether an informal conjecture is a theorem or not, we need to use the particular consequence relation associated with the specific used interpreted language. The problem now is that the specific interpretation is left implicit in conventional discourse, and therefore the needed special consequence relation is generally unknown. We may use the standard consequence relation only if we can ensure that all relevant background knowledge is explicitly added to the theory. However, the acquisition of the relevant background knowledge is a far from trivial task for precisely this knowledge is taken for granted and left implicit in standard scientific discourse.

3 *Logical Analysis*

In order to investigate consequence relations for interpreted languages, we need to make our discussion a bit more precise. This initial formal theory in first-order logic will have a number of axioms, denoted with Σ_{exp} for the explicit axioms of the (initial) theory, and a set of conjectures Γ for the statements that the theory claims to predict or explain. Now let $\models$ denote the standard (tarskian) consequence relation for an uninterpreted first-order language, and let $\models_{\mathsf{theory}}$ denote the unknown non-standard consequence relation of the specific interpreted first-order language of our theory. We want to investigate the logical dependencies between these two possible consequence relations that can be used to determine whether a conjecture $\gamma \in \Gamma$ is derivable from the explicitly mentioned axioms Σ_{exp}. The four logical possibilities in Table 1 present themselves.

Let us first consider case I, $\Sigma_{\mathsf{exp}} \models \gamma$ and $\Sigma_{\mathsf{exp}} \not\models_{\mathsf{theory}} \gamma$. In the case of

[2] For an example of such an interpreted first-order language, see the the language of *Tarski's World* that features prominently in a textbook on logic (Barwise and Etchemendy 1992). We will later draw upon some examples from this book. An interesting discussion of logical consequence relations can be found in (Etchemendy 1990).

	$\Sigma_{\mathsf{exp}} \models_{\mathsf{theory}} \gamma$	$\Sigma_{\mathsf{exp}} \not\models_{\mathsf{theory}} \gamma$
$\Sigma_{\mathsf{exp}} \models \gamma$	II.	I.
$\Sigma_{\mathsf{exp}} \not\models \gamma$	III.	IV.

Table 1: Noninterpreted and Interpreted Consequences.

an interpreted first-order logic, this cannot occur. The non-standard consequence relation $\models_{\mathsf{theory}}$ will be supraclassical: all $\models$-consequences are also $\models_{\mathsf{theory}}$-consequences.[3] Some theorems will hold irrespective of the specific interpretation of the language, that is, they will hold in any interpretation of the language (including the intended interpretation). This gives us the reassurance that we can immediately conclude that $\Sigma_{\mathsf{exp}} \models_{\mathsf{theory}} \gamma$ in case we find that $\Sigma_{\mathsf{exp}} \models \gamma$ (case II).[4] As a result, if we treat an interpreted first-order language as if it were an uninterpreted language, then we can be sure that the theorems we find (using the standard consequence relation $\models$) are also theorems in the interpreted language (using $\models_{\mathsf{theory}}$).

However, as argued above, the used terminology will have antecedent meaning. Therefore, we generally expect that several of the informal conjectures will depend on the specific intended interpretation of the language. So what should we do in case a conjecture is not a $\models$-consequence of the explicit axioms, i.e., when $\Sigma_{\mathsf{exp}} \not\models \gamma$? One option is case IV, the informal conjecture is no theorem, i.e., when also $\Sigma_{\mathsf{exp}} \not\models_{\mathsf{theory}} \gamma$. We will return to case IV below. The remaining option is case III, the informal conjecture is a theorem when we take the interpretation of the language into account, that is when $\Sigma_{\mathsf{exp}} \not\models \gamma$ and $\Sigma_{\mathsf{exp}} \models_{\mathsf{theory}} \gamma$. This is the crucial case for here it would be an important failure to ignore the (implicit) interpretation of the language—we would falsely judge a theorem as a false conjecture. What can we do to prevent this?

The obvious way out is to find a way to ensure that all the relevant implicit background knowledge is explicitly added to the formal theory. Of course, if we would have an axiomatization of all underlying background knowledge, call this set Σ_{imp}, then there would be no more implicit relations between atomic sentences, and we could use the standard consequence relation. In case the background knowledge is first-order expressible (which we may assume in case of an interpreted first-order language) and finitely axiomatizable, we have that

$$\Sigma_{\mathsf{exp}} \models_{\mathsf{theory}} \gamma \text{ if and only if } \Sigma_{\mathsf{exp}} \cup \Sigma_{\mathsf{imp}} \models \gamma$$

Under these conditions, we can reduce the question of how to use the unknown non-standard consequence relation, to the question of how to make the relevant part of the implicit background knowledge explicit. The situation we are interested

[3]This is true for interpreted versions of classical logic. If we consider interpreted non-classical logics, the underlying consequence relation will not satisfy structural properties like monotony, and the resulting logic need not be supraclassical. This points to considerable difficulty in establishing what is implied by an interpreted nonmonotonic theory.

[4]An second result is that, by contraposition, $\Sigma_{\mathsf{exp}} \not\models_{\mathsf{theory}} \gamma$ implies $\Sigma_{\mathsf{exp}} \not\models \gamma$.

in can now be reformulated as:

$$\Sigma_{\mathsf{exp}} \not\models \gamma \text{ and } \Sigma_{\mathsf{exp}} \cup \Sigma_{\mathsf{imp}} \models \gamma$$

with Σ_{imp} being the unknown set of implicit background knowledge. Our goal is now to make relevant parts of Σ_{imp} explicit.[5]

We can push our analysis even further by considering this situation in terms of formal semantics. A first observation is that $\Sigma_{\mathsf{exp}} \not\models \gamma$ implies that there must exist models $\mathcal{M}$ such that $\mathcal{M} \models \Sigma_{\mathsf{exp}} \cup \{\neg\gamma\}$. In fact, constructing such a model would be one of the straightforward ways of proving that $\Sigma_{\mathsf{exp}} \not\models \gamma$. Moreover, there is nothing magical about the construction of these models for it involves only the explicitly known axioms and the conjecture, and the standard consequence relation—a simple algorithm suffices for constructing these models (as we will illustrate in the next section). Each of these models represents a counterexample against the derivation of the conjecture under the assumption that the language is uninterpreted. In our case, however, the conjecture would become derivable in case we would succeed in explicitly adding the background knowledge that enforces the interpreted language, that is $\Sigma_{\mathsf{exp}} \cup \Sigma_{\mathsf{imp}} \models \gamma$. A second observation is that all the models that are counterexamples (in the uninterpreted case) must be violating the implicit background knowledge. That is, for all these models $\mathcal{M}$, it must be the case that $\mathcal{M} \not\models \Sigma_{\mathsf{imp}}$ (since $\mathcal{M} \models \Sigma_{\mathsf{exp}} \cup \{\neg\gamma\}$ and $\mathcal{M} \not\models \Sigma_{\mathsf{exp}} \cup \Sigma_{\mathsf{imp}} \cup \{\neg\gamma\}$). The models that are counterexamples in the uninterpreted case are 'witnesses' of the implicit background knowledge that we need to add explicitly to the axiomatization. Therefore, finding such models can allow us to come to grips with the implicit background knowledge. Consider what happens when we inspect such a model: it necessarily conflicts with some part of our implicit background knowledge on the domain of the theory. To a human observer these models appear strange or extraordinary in some respects. This will prompt us to formulate appropriate axioms that will prevent these models from occurring—axioms that make part of the implicit background knowledge explicit (i.e., some elements of Σ_{imp}). Since these background axioms are based on the specific models that we have examined, this need not be a one-step approach. Further testing may reveal different counterexamples, giving rise to more of the background knowledge being made explicit.

This will, in general, not lead to the axiomatization of all underlying background knowledge. This is hardly unfortunate, since there seems to be no end to the underlying background knowledge. Attempting an axiomatization of all the background knowledge that is taken for granted is at least impractical, if not impossible. We propose to use the informal conjectures for determining which parts of the background knowledge are relevant for the question at hand. That is, we want to use it as a sufficient condition for relevance: if some implicit background knowledge is used for deriving one of the informal conjectures, this is a tell-tale sign for its relevance. In this case there are obvious benefits to making these particular background assumptions explicit, for example, they can become part of future

[5]Note that, even in case the total background knowledge is not first-order expressible or not finitely axiomatizable, some parts of it may still be.

discussion. Note that we do not think that this is a necessary condition, there may be other reasons for including parts of the background knowledge. Also, in a later stage one may want to extend the set of conjectures we want to explain, which may require more of the background knowledge to be explicitly added to the theory.

It is well known that, from a logical point of view, one can always find some additional assumptions that will make a conjecture derivable (Quine 1953, p.43). So it is a legitimate concern if we are able to distinguish false conjectures from informal conjectures that can be made derivable by explicating background knowledge (case IV in Table 1 we delayed discussing above). That is, how can we identify false conjectures, i.e., conjectures for which we have that $\Sigma_{\mathsf{exp}} \not\models_{\mathsf{theory}} \gamma$? Inspection of the models that are counterexamples provides an easy safe-guard against this. In this case there will, again, be models $\mathcal{M}$ such that $\mathcal{M} \models \Sigma_{\mathsf{exp}} \cup \{\neg\gamma\}$. However, since in this case $\Sigma_{\mathsf{exp}} \cup \Sigma_{\mathsf{imp}} \not\models \gamma$, some of these counterexamples will be in perfect harmony with all the background knowledge that we would take for granted, i.e., $\mathcal{M} \models \Sigma_{\mathsf{exp}} \cup \Sigma_{\mathsf{imp}} \cup \{\neg\gamma\}$. Inspection of these models will reveal a genuine counterexample—an intended model of the theory in which the conjecture fails—proving that the informal conjecture does not hold. We may only rebut a potential counterexample by relying on unproblematic background knowledge. Otherwise, we must conclude that $\Sigma_{\mathsf{exp}} \not\models_{\mathsf{theory}} \gamma$.

4 *Applications and Computational Support*

Our discussion up to this point has been rather abstract. However, as we will show in this section, our analysis above can be directly applied to concrete situations. In particular, we will show how this can immediately be supported by standard tools from automated reasoning. The formalization of an empirical science theory is typically using an interpreted language, with the interpretation being enforced by the implicit background knowledge that is taken for granted. Barwise and Etchemendy (1992) introduce the interpreted first-order language of Tarski's world with an associated computer program that visualizes this blocks world. The vocabulary of this first-order language contains constants (a through f plus n1, n2, ...), predicates (unary predicates: Tet, Cube, Dodec, Small, Medium, Large; binary predicates: $=$, Smaller, Larger, LeftOf, RightOf, BackOf, FrontOf; and a tertiary predicate Between), and no functions. The predicates have a fixed interpretation in the associated computer program, for example an object cannot be both a cube and a tetrahedron. The fixed interpretation assigned by Tarski's world is one that is "reasonably consistent with the corresponding English verb phrase" (Barwise and Etchemendy 1992, p.11). The authors assume that readers share common background knowledge on names of these predicates, and that the program's interpretation is consistent with it. Although the predicates have a very precise meaning, the authors do not give the axioms that are assumed to hold. Instead, they invite the reader to experiment with the program, and get acquainted with their meaning by trail and error—not unlike in ordinary language acquisition. We can use some examples from (Barwise and Etchemendy 1992) in order to illustrate the strategy for elucidating implicit background knowledge discussed in the previous section.

4.1 *Interpreted consequence*

We are particularly interested in arguments which depend on the fact that the language of Tarski's world is an interpreted language. In this case, the (formal) proofs do strictly depend on the interpretation as given in the program, and would not hold for arbitrary interpretations of the predicates. If we would substitute the used predicate symbols with fresh ones having the same associated arity, the arguments would not hold. An exercise which relies on the specific interpretation of the predicates is (Barwise and Etchemendy 1992, Problem 5-30, p.143).

Is $\exists x\ [\mathsf{FrontOf}(\mathsf{c}, x) \wedge \mathsf{Cube}(x)]$ a consequence of

A1. $\forall x\ [\mathsf{Cube}(x) \vee (\mathsf{Tet}(x) \wedge \mathsf{Small}(x))]$
A2. $\exists x\ [\mathsf{Large}(x) \wedge \mathsf{BackOf}(x, \mathsf{c})]$

According to the instructor's manual this is indeed the case in Tarski's world (Eberle 1993). It will be impossible to build a world that is a counterexample using the Tarski's World program. Is it also valid for arbitrary interpretations of the predicates? To answer this question, we can use standard tools from automated reasoning, like automated theorem prover OTTER (McCune 1994b) and automated model generator MACE (McCune 1994a). The answer turns out to be negative: theorem prover OTTER fails to find a proof, and model generator MACE has no trouble in finding counterexamples. These counterexamples are models of the premises in which the conjecture is false (the first model on universe $\{0, 1\}$ is shown in Table 2).[6]

c	
	0

Cube	
0	T
1	T

Tet	
0	F
1	F

Small	
0	F
1	F

BackOf	0	1
0	T	F
1	F	F

FrontOf	0	1
0	F	F
1	F	F

Large	
0	T
1	F

Table 2: Counterexample I.

[6] For example, by invoking MACE with the options '`-n2 -p -m100`' (see for details McCune 1994a). MACE generates 16 models on $\{0, 1\}$ in less than a second. A formal model consists of a universe (here the two elements $\{0, 1\}$) and a mapping between the non-logical symbols (here constant c; unary predicates Cube, Tet, Small, and Large; and binary relations BackOf and FrontOf) and elements of the universe. For example, consider the model in Table 2: here, the constant symbol c is interpreted as object 0, and predicate symbol Cube is interpreted as both $\mathsf{Cube}(0)$ and $\mathsf{Cube}(1)$ are true. That is, all objects in the universe of this model are cubes making the sentence $\forall x\ [\mathsf{Cube}(x) \vee (\mathsf{Tet}(x) \wedge \mathsf{Small}(x))]$, the first premise, is indeed true in this model.

Finding this model proves that the argument does not hold using a standard consequence relation, in symbols,

$$\{\mathrm{A1}, \mathrm{A2}\} \not\models \exists x\, [\mathsf{FrontOf}(\mathsf{c}, x) \wedge \mathsf{Cube}(x)]$$

However, the argument should hold when we respect the interpretation of the predicates. According to our above discussion, the counterexample in Table 2 must conflict with the interpretation of the predicates in the Tarski's world program. Moreover, this model must also be in conflict with the ordinary language meaning of the corresponding English phrases. That is, anybody with some proficiency in English should find this model in violation of his or her background knowledge of the domain. Our expectation is that, when confronted with this model, a person is able to articulate *why* this model should not be allowed to occur.

Upon inspecting the model in Table 2, we immediately note a strange feature: $\mathsf{BackOf}(0, 0)$ is true, the object 0 is in back of itself. This is not in accordance with the normal English interpretation of this predicate, and we decide to spell out this background knowledge explicitly:

B1. $\forall x\, [\neg\mathsf{BackOf}(x, x)]$

After adding this background assumption explicitly to the premises, the model in Table 2 will no longer be a model of the theory. We can now test anew if we can now formally derive the conclusion. Notice that this need not be the case for there may exist different counterexamples. Indeed, theorem prover OTTER still fails to find a proof, and model generator MACE is able to construct further counterexamples (now 8 on $\{0, 1\}$, the first is shown in Table 3).

c	1

Cube	
0	T
1	T

Tet	
0	F
1	F

Small	
0	F
1	F

BackOf	0	1
0	F	T
1	F	F

FrontOf	0	1
0	F	F
1	F	F

Large	
0	T
1	F

Table 3: Counterexample II.

There must still be more background knowledge at stake. Inspecting the model in Table 3, there seems to be no problem with the interpretation of each predicate independently. However, some natural relations between the predicates are not properly taken into account. In the model $\mathsf{BackOf}(0, 1)$ is true while at the same time $\mathsf{FrontOf}(1, 0)$ is false. This conflict with our background understanding of these predicates. Our intuitions say that these two predicates are inversely related, so we decide to explicitly add this relation between FrontOf and BackOf:

B2. $\forall x, y\, [\mathsf{FrontOf}(x, y) \leftrightarrow \mathsf{BackOf}(y, x)]$

Have we now added all relevant background knowledge? We test again using theorem prover OTTER, but still fail to find a proof. Yet again, the model generator MACE is able to construct further counterexamples (still 4 on $\{0, 1\}$, the first is shown in Table 4).

c	
	0

Cube	
0	T
1	F

Tet	
0	F
1	T

Small	
0	F
1	T

BackOf	0	1
0	F	F
1	T	F

FrontOf	0	1
0	F	T
1	F	F

Large	
0	F
1	T

Table 4: Counterexample III.

Again, we examine this new counterexample to verify whether it is an intended model of this domain. Inspection reveals that this model is not conform the normal English interpretation of the predicates: both $\mathsf{Small}(1)$ and $\mathsf{Large}(1)$ are true, the same object is both small and large. We do not want to exclude that an object is neither small or large (as object 0 in Table 4) for, after all, there might be medium sized objects. An object being both small and large at the same time, however, conflicts with our implicit understanding of these two predicates, and we decide to add a further background assumption explicitly to the theory:

B3. $\forall x \neg[\mathsf{Small}(x) \ \wedge \ \mathsf{Large}(y)]$

Did we now make all relevant background knowledge explicit? At last, the answer is positive: theorem prover OTTER finds a proof using the two premises and two of the background assumptions (B2 and B3).[7] The proof constructed by OTTER is a clause-based resolution proof. Paraphrasing this formal proof, we find that OTTER derives from premise A2 and background assumption B2, that c is in front of a large object; and using premise A1 and background assumption B3, that this large object must be a cube. That is, we can now formally derive that c is in front of a cube, in symbols,

$$\{\mathrm{A1}, \mathrm{A2}, \mathrm{B2}, \mathrm{B3}\} \models \exists x \, [\mathsf{FrontOf}(\mathsf{c}, x) \wedge \mathsf{Cube}(x)]$$

The answer to this problem is, indeed, positive. We have now proved that the argument is valid when respecting the (implicit) interpretations of Tarski's world, in symbols,

$$\{\mathrm{A1}, \mathrm{A2}\} \models_{\mathsf{TW}} \exists x \, [\mathsf{FrontOf}(\mathsf{c}, x) \wedge \mathsf{Cube}(x)]$$

[7] That is, we may decide to relax the first background assumption B1 again because it is not necessary for this argument. Notice that this implies that the model in Table 2 is also violating the other background assumptions, otherwise this counterexample would still disprove the argument (in this case B2).

4.2 *Interpreted non-consequence*

An exercise with the same premises as above is (Barwise and Etchemendy 1992, Problem 5-31, p.143).

> Assume the following premises:
>
> A1. $\forall x\ [\mathsf{Cube}(x) \vee (\mathsf{Tet}(x) \wedge \mathsf{Small}(x))]$
> A2. $\exists x\ [\mathsf{Large}(x) \wedge \mathsf{BackOf}(x, \mathsf{c})]$
>
> Does it follow that $\neg\exists x\ [\mathsf{Small}(x) \wedge \mathsf{BackOf}(x, \mathsf{c})]$?

We are asked to establish whether this argument is valid when respecting the interpretation of predicates. According to the instructor's manual this is not the case in Tarski's world (Eberle 1993). Since we established above that B1, B2, and B3 are part of the implicit background knowledge, we will start with the set of explicit premises $\{\mathrm{A1}, \mathrm{A2}, \mathrm{B1}, \mathrm{B2}, \mathrm{B3}\}$. As expected, theorem prover OTTER fails to derive the conjecture from this set of premises. We resort to model generator MACE in order to find models of the premises in which the conjecture is false, that is, in which there exists a small object in the back of object c. Model generator MACE fails to find any model of cardinality 2, but produces 24 models of cardinality 3 (the first of them is reproduced in Table 5).

c	
	1

Cube	
0	T
1	T
2	T

Tet	
0	F
1	F
2	F

Small	
0	F
1	F
2	T

BackOf	0	1	2
0	F	T	F
1	F	F	F
2	F	T	F

FrontOf	0	1	2
0	F	F	F
1	T	F	T
2	F	F	F

Large	
0	T
1	F
2	F

Table 5: Counterexample IV.

Can we rebut this model by mobilizing part of the background knowledge? Examining the model in Table 5, we have a configuration of one cube c that is placed in front of two other cubes, one of which is large (as required by premise A2), and the other small (refuting the conjecture). The interpretation of the predicates in this model is conform our implicit background knowledge. We can confirm this by replicating a corresponding world using the Tarski's World program. We must conclude that this model is a genuine counterexample disproving the conjecture. That is, we have proved that the argument does not hold when respecting the interpretations of Tarski's world, in symbols,

$$\{\mathrm{A1}, \mathrm{A2}\} \not\models_{\mathsf{TW}} \neg\exists x\ [\mathsf{Small}(x) \wedge \mathsf{BackOf}(x, \mathsf{c})]$$

These simple examples demonstrate the necessity of taking implicit background knowledge into account in languages that have antecedent meaning. Moreover, they illustrate how automated reasoning tools can assist in the acquisition of implicit background knowledge by constructing the models that are in conflict with our understanding of the domain. This has also proved to be crucial in more substantial applications: uncovering implicit theoretical presuppositions is one of the main problems in the reconstruction of informal sociological theories (Kamps and Pólos 1999; Kamps 1999). A word of warning is in place for it is important not to underestimate the general complexity of this task. The use of automated tools is subject to important limitations, both in principle (first-order logic is not decidable), as in practice (time, memory, CPU-power). The above examples are well within these limits: none of the successful or failed proof attempts or model searches lasted more than a single second. It is interesting to note that the models conflicting with our implicit understanding of the domain are particularly difficult to find by hand. Since we ourselves possess the underlying background knowledge, we have a natural tendency to focus our attention towards the intended models of the theory. Computer programs, not endowed with this underlying understanding, are not hindered with such a bias.

5 *Discussion and Conclusions*

Scientific theories about empirical phenomena are human constructions. Formal theories do interface, on the one hand, with empirical reality, leading to all the familiar problems of confirmation, falsification, and truth approximation. On the other hand and less frequently discussed, formal theories also interface with human conceptions and theoretical intuitions. It is well-known that in the empirical sciences, terms "have a clear meaning independent of the theory and they retain this meaning within the context of the theory" (Kuipers 2001, p.44). The terminology is "antecedently understood" (Hempel 1966, p.75). One of the main reasons for the formalization of scientific theories is to bring out the meaning of concepts in an explicit fashion (Suppes 1968). Making the underlying background knowledge explicit contributes to our understanding of the theory by avoiding ambiguity.[8] This may work even if there is no full consensus on the meaning of terminology (as is rarely the case in the social sciences). In case of partial consensus, researchers would still agree that some background axioms should hold. At the same time, making the underlying background knowledge explicit is a highly non-trivial task. This is immediately clear once we realize that much of our background knowledge is tacit knowledge (Polanyi 1958). This means that, even though we are carriers

[8]Arguably, it is more acceptable to revise implicit background knowledge, than to retract some explicit statements of a theory. That does not imply that explicit background knowledge cannot evolve over time. In fact, formalization is known to trigger the further development of terminology. Even in the simplified setting of Tarski's World the background knowledge may change over time: in the new version of the program, the interpretation of the Between predicate has changed (Barwise and Etchemendy 1999).

of implicit background knowledge, the articulation of it may be beyond our own control. That raises the question whether we can ever be sure that all relevant background knowledge has been made explicit. If we must assume, as Polanyi does, that part of our background knowledge will always remain implicit, this has some important methodological implications.

The general conclusion is a call for caution when discussing what are the sets of consequences or models of a formal theory. If we cannot be sure that all background knowledge is explicitly added to the theory, we must anticipate that we can only derive part of its consequences, and that the set of formal models of the theory contains models that are conflicting with our intuitions. In particular, this may interfere with attempts to compare formal theories by their sets of consequences or models, as in approaches to truth-likeness (Kuipers 2000). If we are to compare theories by the statements they imply, we must take into account that we are systematically underestimating the set of consequences. Hence, in general, a statement approach to truthlikeness will miss out some of the successes and failures of a theory. If we compare theories by their models, we must take into account that we are systematically overestimating the set of models. Thus a semantic approach to truthlikeness will find, in general, more successes and failures than warranted by the theory. Keeping this in mind, there is even more reason to pursue efforts to make relevant parts of the implicit background knowledge explicit. After all, if our implicit background knowledge is (approximately) true then adding this explicitly to a formal theory should bring us even closer to the truth.[9]

Background knowledge not only affects the hypothetical problem of enumerating deductively closed sets of consequences or all the models of a theory. Even apart from the question of implicit background knowledge, comparing all consequences or models of a theory is already infeasible in practice for these sets are generally infinite. As a result, theory comparison is relativized to a particular set of key predictions (such as the comparison of electrodynamic theories shown in [Kuipers 2001, Table 8.2, p.236]). Precisely in such a setting we would want to avoid falsely discarding conjectures by not fully taking into account the meaning of terminology. In our discussion of the axiomatization in formal logic above, we have focused on this case by assuming that a specific conjecture is at stake. We have shown how formal semantics may be used to avoid the unjustified rejection of a conjecture. Just as logic provides a formal notion of proof, it also provides a formal notion of refutation. A formal refutation of a conjecture is a formal model (in the logical sense) of the premises in which the conjecture does not hold. In

[9]Technically this will be somewhat more involved, since it points out an asymmetry between the statement and models view on a theory. The 'extra' models of the theory are necessarily all in conflict with the implicit background knowledge (i.e., these are all nonintented or nonsensical models). Thus, we will approximate truth in terms of models. However, even if the background knowledge is (approximately) true, the 'missed' consequences of the theory may contain both interesting statements and nonsensical ones. As a result, explicitly adding background knowledge may increase both the number of successes and failures of the theory (in terms of statements). Only in case the explicit axioms of the theory are also (approximately) true, the 'missed' consequences will also be all (approximately) true. Then, we will also approximate truth in terms of consequences of the theory.

the context of implicit background knowledge, a formal refutation need not correspond to an empirical refutation (only models that respect the terminology may correspond to an empirical possibility). If we are able to construct a formal model refuting a conjecture, we can inspect the model to verify whether the refutation is warranted. If this is not the case, inspecting the model immediately suggests which background knowledge needs to be added explicitly to the theory. Recall that much of our underlying knowledge is tacit, however, this need not prevent us from identifying models that are in conflict with it. Identifying such a model makes us aware of our tacit understanding, and can provide crucial help in its articulation.[10] This results in an interesting interplay between conjectures, proofs, and refutations. Although the antecedent meaning of terminology is a principal feature of the empirical sciences, we may also have background knowledge on terminology in non-empirical fields like mathematics and philosophy. In fact, our discussion shows some remarkable similarities with discussions on mathematical discovery (Pólya 1945; Lakatos 1976).[11] The main difference is that in the non-empirical sciences, we have the luxury of being able to stipulate that the concept as characterized in a theory is the 'real' concept.

It is known for long that logical axiomatization can contribute to theory development in the empirical sciences (Woodger 1937; Kyburg 1968). In recent years, this has resulted in the formalization of a number of sociological theories (Péli et al. 1994; Hannan 1998; Kamps and Pólos 1999). Having this is mind, we find it difficult to agree with the remark that a logical axiomatization or so-called statement approach is not very useful and very difficult (compared with a semantic approach). Kuipers (2001, p.319) has it that

> the statement approach is certainly more difficult for specific reconstructions. ... Happily enough, not all interesting theoretical questions need logical treatment. ... Given our intention to be as useful as possible for actual scientific research we will restrict our attention to the structuralist approach.

We do not disagree on the merits of semantic approaches. There are many examples of fruitful axiomatization in the structuralist approach (Balzer et al. 1987, 2000). We also immediately admit that, more generally, a semantic approach has specific advantages over a statement approach. To mention just a few, a semantical approach immediately suggests itself for establishing the consistency of a theory or domain, or for disproving a conjecture. However, we disagree on the decision to de-emphasize logical axiomatizations. Generally speaking, a statement approach

[10] It is important to bear in mind that tacit knowledge can be made explicit, and that doing so has contributed to the theoretical development of various fields (Polanyi 1958). This does, however, require significant effort, and it will be impossible to formalize all the tacit knowledge in a particular field—a point with which we concur.

[11] Our discussion of models conflicting with the implicit antecedent meaning of terminology shows resemblance with Lakatos' monster-barring heuristic (dealing with doughnut-shaped or picture-frame polyhedra discussed in [Pólya 1954, p.42] and [Lakatos 1976, p.19]).

also has specific merits, think of establishing the inconsistency of a theory. For certain cases a statement approach is intuitively more appropriate. It is of interest to analyze reasons that might explain this discrepancy between these views on formal theorizing.[12]

Perhaps a difference in appreciation of the statement and semantical approaches is rooted on the difference between the sciences. Our experience in logical reconstruction has focused on informal theories in sociology, whereas the structuralist approach is based on reconstructions of mathematical physics (Sneed 1971), although later also applied to various other fields (Balzer et al. 1987), including sociology (Manhart 1994). The axiomatization of a highly mathematical theory would also require the axiomatization of the used mathematical techniques. This is a highly nontrivial task in case of the advanced, quantitative mathematics used in mathematical physics. This view is consistent with the axiomatization of one of the rare mathematical theories in sociology, a mathematical model of social groups (Simon 1952). The resulting axiomatization is almost completely concerned with the differential equations used in the mathematical model (Kyburg 1968, Ch.12). The structuralist approach, in contrast, allows for freely using all kinds of useful mathematics, allowing the reconstruction to focus on the theory at hand without first having to axiomatize various mathematical theories. This will make reconstructions certainly easier in case of advanced mathematical theories such as in theoretical physics. However, the mathematical finesse of physics is not a rule in the empirical sciences. In fact, in fields like sociology, mathematical theories are even rare, and the standard discourse is in natural language. At least for non-mathematical theories in the empirical sciences, the statement approach to formalization seems a viable option.

It is important to note that the flexibility of the structuralist approach does not come without a price. Since the standard mathematical vernacular is only partially formal, it requires substantial mathematical background knowledge usually shaped by years of mathematical training. If the goal is to provide a computational implementation of a theory, we are again confronted with the fact that computers lack the mathematical background knowledge. It is unclear to what extent a structuralist formalization renders our theories in a form that can be interpreted by a computer.[13] Standard mathematics is usually too informal to allow for construct-

[12]Some have argued that there is some form of resentment against logical empiricism (Friedman 1991, 1999). There may be some truth in this, e.g., the structuralist approach is also sometimes referred to as the "non-statement view" (Stegmüller 1973). Needless to say, the field of logic has changed dramatically since the days of positivism. As Hintikka (1998, p.304) writes: "[W]hen the sharpest philosophers of science realized that a study of 'the logical syntax of the language of science' was not enough, they resorted to set theory for their conceptualizations. Ironically some misguided philosophers of science have continued to seek salvation in set theory long after the development of logical semantics and systematic model theory." However interesting such arguments may be from a historical point of view, we will restrict ourselves here to substantial reasons.

[13]Although Kuipers (2001, p.302) writes: "it will become quite clear ... that the structuralist analysis of theories can almost directly be used for the computational representation of theories." This is far from obvious to me, in fact, it seems to require pencil, paper, and a philosophy professor

ing formal proofs. Formal logic, in contrast, provides the needed rigorousness. A formalization using the so-called statement approach immediately allows for computational implementation. In fact, the automation of logical reasoning is one of the oldest applications of artificial intelligence (Newell and Simon 1956; Beth 1958). Current implementations of automated reasoning programs are powerful tools that can support the formal reconstruction of theories in various ways (Kamps 1998, 1999). By using such tools, the construction of a logical axiomatization need not be more difficult than a structuralist reconstruction. One of the reasons why logical axiomatization is considered to be difficult, is because manually deriving theorems using a particular formal proof system can be painstaking and prone to errors. Unlike humans, computers are well-equipped for performing tedious tasks like proof checking or proof finding in a formal proof system. In fact, the detailed rigorousness is precisely what makes a logical axiomatization suitable for computational reasoning. In sum, using these programs can greatly facilitate the process of reconstructing scientific theories in formal logic. Moreover, if the aim is to provide a computational representation of theories, an axiomatization produced by the statement approach is still an attractive alternative.

In our experience, both a statement approach and a semantical approach have their respective merits. Since these merits do not coincide, it is of particular interest to investigate ways that can exploit both views. In this light, it is important to note that most logics come with both a proof theory and a formal semantics. This allows us to view our theory as either a set of statements and a set of models, depending on which point of view is better suited for the question at hand. For example, for proving a particular conjecture we can use syntactic proof theory and for disproving a particular conjecture we can use semantic model theory. That is, the "pragmatic choice" between a "statement approach" and a "semantic approach" as discussed in (Kuipers 2001, p.319) need not be made: there is no reason why we cannot have the best of both worlds. Our earlier discussion on background knowledge is an illustrative example of how we can exploit a semantical view on the statement approach. One can only hope that such considerations may ultimately lead to a reconciliation of the two approaches.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grant # 400-20-036).

Jaap Kamps
Institute for Logic, Language and Computation
University of Amsterdam
Nieuwe Achtergracht 166
NL-1018WV Amsterdam
the Netherlands
Email: ⟨kamps@illc.uva.nl⟩

in order to operate a structuralist representation.

REFERENCES

W. Balzer, C. U. Moulines, and J. D. Sneed. *An Architectonic for Science: The Structuralist Program*, volume 186 of *Synthese Library*. D. Reidel Publishing Company, Dordrecht, 1987.

W. Balzer, C. U. Moulines, and J. D. Sneed, editors. *Structuralist Knowledge Representation: Paradigmatic Examples*, volume 75 of *Poznań Studies in the Philosophy of the Sciences and the Humanities*, 2000. Rodopi, Amsterdam.

J. Barwise and J. Etchemendy. *The Language of First-Order Logic*. CSLI Publications, Stanford CA, revised and expanded third edition, 1992.

J. Barwise and J. Etchemendy. *Language, Proof and Logic*. Seven Bridges Press, New York NY and CSLI Publications, Stanford CA, 1999.

E. W. Beth. On machines which prove theorems. *Simon Stevin: Wis- en Natuurkundig Tijdschrift*, 32:49–60, 1958.

R. Eberle. *Instructor's Manual to accompany Jon Barwise and John Etchemendy's The Language of First-Order Logic*. CSLI Publications, Stanford University, 1993.

H. B. Enderton. *A Mathematical Introduction to Logic*. Academic Press, New York, 1972.

J. Etchemendy. *The Concept of Logical Consequence*. Harvard University Press, Cambridge, 1990.

M. Friedman. The re-evaluation of logical positivism. *The Journal of Philosophy*, LXXXVIII:505–519, 1991.

M. Friedman. *Reconsidering Logical Positivism*. Cambridge University Press, Cambridge UK, 1999.

M. T. Hannan. Rethinking age dependence in organizational mortality: Logical formalizations. *American Journal of Sociology*, 104:126–164, 1998.

C. G. Hempel. *Philosophy of Natural Science*. Foundations of Philosophy Series. Prentice-Hall, Englewood Cliffs NJ, 1966.

J. Hintikka. Truth definitions, skolem functions and axiomatic set theory. *The Bulletin of Symbolic Logic*, 4:303–337, 1998.

J. Kamps. Formal theory building using automated reasoning tools. In A. G. Cohn, L. K. Schubert, and S. C. Shapiro, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR'98)*, pages 478–487. Morgan Kaufmann Publishers, San Francisco CA, 1998.

J. Kamps. On criteria for formal theory building: Applying logic and automated reasoning tools to the social sciences. In J. Hendler and D. Subramanian, editors, *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 285–290. AAAI Press/The MIT Press, Menlo Park CA, 1999.

J. Kamps and L. Pólos. Reducing uncertainty: A formal theory of *Organizations in Action*. *American Journal of Sociology*, 104:1776–1812, 1999.

T. A. F. Kuipers. *From Instrumentalism to Constructive Realism: On Some Relations between Confirmation, Empirical Progress, and Truth Approximation*, volume 287 of *Synthese Library*. Kluwer Academic Publishers, Dordrecht, 2000.

T. A. F. Kuipers. *Structures in Science: Heuristic Patterns Based on Cognitive Structures*, volume 301 of *Synthese Library*. Kluwer Academic Publishers, Dordrecht, 2001.

H. E. Kyburg, Jr. *Philosophy of Science: A Formal Approach*. The Macmillan Company, New York, 1968.

I. Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, Cambridge, England, 1976.

K. Manhart. Strukturalistische Theorienkonzeption in den Sozialwissenschaften. *Zeitschrift für Soziologie*, 23:111–128, 1994.

W. McCune. A Davis-Putnam program and its application to finite first-order model search: Quasigroup existence problems. Technical report, Argonne National Laboratory, Argonne IL, 1994a. DRAFT.

W. McCune. OTTER: Reference manual and guide. Technical Report ANL-94/6, Argonne National Laboratory, Argonne IL, 1994b.

A. Newell and H. A. Simon. The logic theory machine. In *IRE Transactions on Information Theory*, volume IT-2(3), pages 61–79, 1956.

G. Péli, J. Bruggeman, M. Masuch, and B. Ó Nualláin. A logical approach to formalizing organizational ecology. *American Sociological Review*, 59:571–593, 1994.

M. Polanyi. *Personal Knowledge: Towards a Post-Critical Philosophy*. University of Chicago Press, Chicago IL, 1958.

G. Pólya. *How to Solve It: a new aspect of mathematical method*. Princeton University Press, Princeton New Jersey, 1945.

G. Pólya. *Induction and Analogy in Mathematics*, volume I of *Mathematics and Plausible Reasoning*. Princeton University Press, Princeton New Jersey, 1954.

K. R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge and Kegan, London, 1963.

W. V. O. Quine. *From a logical point of view*. Harvard University Press, Cambridge MA, 1953.

H. A. Simon. A formal model of interaction in social groups. *American Sociological Review*, 17:202–211, 1952.

J. D. Sneed. *The Logical Structure of Mathematical Physics*. D. Reidel Publishing Company, Dordrecht, 1971.

W. Stegmüller. *Logische Analyse der Struktur ausgereifter physikalischer Theorien, 'Non-statement view' von Theorien*. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie: Band II Theorie und Erfahrung, Teil D. Springer-Verlag, Berlin, 1973.

P. Suppes. The desirability of formalization in science. *Journal of Philosophy*, LXV(20): 651–664, 1968.

A. Tarski. *Introduction to Logic and to the Methodology of Deductive Sciences*. Oxford University Press, New York, second, revised edition, 1946.

J. H. Woodger. *The Axiomatic Method in Biology*. Cambridge University Press, 1937.