On Criteria for Formal Theory Building: Applying Logic and Automated Reasoning Tools to the Social Sciences

Jaap Kamps
Applied Logic Laboratory
Institute for Logic, Language and Computation
University of Amsterdam
Sarphatistraat 143, 1018 GD Amsterdam, the Netherlands
kamps@ccsom.uva.nl

Abstract

This paper provides practical operationalizations of criteria for evaluating scientific theories, such as the consistency and falsifiability of theories and the soundness of inferences, that take into account definitions. The precise formulation of these criteria is tailored to the use of automated theorem provers and automated model generators—generic tools from the field of automated reasoning. The use of these criteria is illustrated by applying them to a first order logic representation of a classic organization theory, Thompson's Organizations in Action.

Introduction

Philosophy of science's classical conception of scientific theories is based on the axiomatization of theories in (first order) logic. In such an axiomatization, the theory's predictions can be derived as theorems by the inference rules of the logic. In practice, only very few theories from the empirical sciences have been formalized in first order logic. One of the reasons is that the calculations involved in formalizing scientific theories quickly defy manual processing. The availability of automated reasoning tools allows us to transcend these limitations. In the social sciences, this has led to renewed interest in the axiomatization of scientific theories (Péli *et al.* 1994; Péli & Masuch 1997; Péli 1997; Bruggeman 1997; Hannan 1998; Kamps & Pólos 1999). These authors present first order logic versions of heretofore non-formal scientific theories.

The social sciences are renowned for the richness of their vocabulary (one of the most noticeable differences with theories in other sciences). Social science theories are usually stated using many related concepts that have subtle differences in meaning. As a result, a formal rendition of a social science theory will use a large vocabulary. We recently started to experiment with the use of definitions as a means to combine a rich vocabulary with a small number of primitive terms.

> Now definitions are unlike theorems and unlike axioms. Unlike theorems, definitions are not things we prove. We just declare them by fiat. But unlike axioms, we do not expect definitions to add substantive information. A definition is expected to add to our convenience, not to our knowledge. (Enderton 1972, p.154)

If dependencies between different concepts are made explicit, we may be able to define some concepts in terms of other concepts, or in terms of a smaller number of primitive concepts. If the theory contains definitions, the defined concepts can be eliminated from the theory by expanding the definitions. Eliminating the defined concepts does not affect the theory, in the sense that the models and theorems of the theory remain the same.

This paper provides practical operationalizations of criteria for evaluating scientific theories, such as the consistency and falsifiability of theories and the soundness of inferences. In earlier discussion of the criteria for evaluating theories, we did not distinguish between different types of premises (Kamps 1998). In this paper, we will provide practical operationalizations of these criteria that take definitions into account, and illustrate their use on a formal fragment of organization theory.

Logical formalization

Most social science theories are stated in ordinary language (except, of course, for mathematical theories in economics). The main obstacle for the formalization of such a discursive theory is their rational reconstruction: interpreting the text, distinguishing important claims and argumentation from other parts of the text, and reconstructing the argumentation. This reconstruction is seldom a straightforward process, although there are some useful guidelines (Fisher 1988). When the theoretical statements are singled-out, they can be formulated in first order logic. The main benefit of the formalization of theories in logic is that it provides clarity by providing an unambiguous exposition of the theory (Suppes 1968). Moreover, the fields of logic and philosophy of science have provided a number of criteria for evaluating formal theories, such as the consistency and falsifiability of theories and the soundness of inferences. Our aim is to develop support for the axiomatization of theories in first order logic by giving specific operationalizations of these criteria. These specific formulations are chosen such that the criteria can be established in practice with relative ease, i.e., such that existing automated reasoning tools can be used for this purpose.[1]

[1] Of course, we hope that this will be regarded as an original contribution, but the claim to originality is a difficult one to establish. The novelty is in the combination of ideas from various fields and our debts to the fields of logic and philosophy of science fan out much further than specific citations indicate.

Criteria for Evaluating Theories

We will use the following notation. Let Σ denote the set of premises of a theory. A formula φ is a theorem of this theory if and only if it is a logical consequence, i.e., if and only if $\Sigma \models \varphi$. The theory itself is the set of all theorems, in symbols, $\{\varphi \mid \Sigma \models \varphi\}$.

Consistency The first and foremost criterion is *consistency*: we can tell whether a theory in logic is contradiction-free. If a theory is inconsistent, it cannot correspond to its intended domain of application. Therefore, empirical testing should focus on identifying those premises that do not hold in its domain. The formal theory can suggest which assumptions are problematic by identifying (minimal) inconsistent subsets of the premises.

The theory is *consistent* if we can find a model $\mathcal{A}$ such that the premises are satisfied: $\mathcal{A} \models \Sigma$. A theory is *inconsistent* if we can derive a contradiction, $\perp$, from the premises: $\Sigma \vdash \perp$.

Soundness Another criterion is *soundness* of arguments: we can tell whether a claim undeniably follows from the given premises. If the derivation of a claim is unsound, empirical testing of the premises does not provide conclusive support for the claim. Conversely, empirically refuting the claim may have no further consequences for the theory. Many of our basic propositions are inaccessible for direct empirical testing. Such propositions can be indirectly tested by their testable implications (Hempel 1966). In case of unsound argumentation, examining the counterexamples provides useful guidance for revision of the theory.

A theorem φ is *sound* if it can be derived from the premises $\Sigma \vdash \varphi$. A theorem φ is *unsound* (i.e., φ is no theorem) if we can construct a counterexample, that is, a model $\mathcal{A}$ in which the premises hold, and the theorem is false: $\mathcal{A} \models \Sigma$ and $\mathcal{A} \not\models \varphi$.

Falsifiability Falsifiability of a theorem means that it is possible to refute the theorem. Self-contained or tautological statements are unfalsifiable—their truth does not depend on the empirical assumptions of the theory. Falsifiability is an essential property of scientific theories (Popper 1959). If no state of affairs can falsify a theory, empirical testing can only reassert its trivial validity. A theory is falsifiable if it contains at least one falsifiable theorem.

An initial operationalization of falsifiability is: a theorem φ is *unfalsifiable* if it can be derived from an empty set of premises: $\vdash \varphi$ and *falsifiable* if we can construct a model $\mathcal{A}$ (of the language) in which the theorem is false: $\mathcal{A} \not\models \varphi$. Note that we cannot require this model to be a model of the theory. A theorem is necessarily true in all models of the theory (otherwise it would not be a theorem). We should therefore ignore the axioms of the theory, and consider arbitrary models of the language. This initial formulation works for some unfalsifiable statements like tautologies, but may fail in the context of definitions. Consider the following simple example: a theory that contains a definition of a $\mathsf{Bachelor}$ predicate.

$$\forall x\ [\mathsf{Bachelor}(x) \ \leftrightarrow\ \mathsf{Man}(x) \wedge \mathsf{Unmarried}(x)]$$

In such a theory, we will have a the following theorem.

$$\forall x\ [\mathsf{Bachelor}(x) \ \rightarrow\ \mathsf{Man}(x)]$$

Using falsifiability as formulated above, we would conclude that this statement is falsifiable. It is easy to construct models (of the language) in which the theorem is false, that is, models in which an object a occurs such that $\mathsf{Bachelor}(a)$ is assigned *true* and $\mathsf{Man}(a)$ is *false*. However, if we expand the definition, then the theorem becomes

$$\forall x\ [\mathsf{Man}(x) \wedge \mathsf{Unmarried}(x) \ \rightarrow\ \mathsf{Man}(x)]$$

This expanded theorem is tautologically true and therefore unfalsifiable by the above formulation. This is problematic, since "we do not expect definitions to add substantive information" (Enderton 1972, see the above quotation). Therefore, the elimination of defined concepts should also not affect any of the criteria for evaluating theories. Although we have to ignore the axioms, we must take the definitions into account when establishing falsifiability.

In the context of definitions $\Sigma_{\mathsf{def}} \subseteq \Sigma$, a theorem φ is *falsifiable* if there exists a model $\mathcal{A}$ in which the definitions hold, and the theorem is false: $\mathcal{A} \models \Sigma_{\mathsf{def}}$ and $\mathcal{A} \not\models \varphi$. A theorem φ is *unfalsifiable* if the theorem can be derived from only the set of definitions: $\Sigma_{\mathsf{def}} \vdash \varphi$.

Satisfiability Satisfiability is the counterpart of falsifiability. Satisfiability of a theorem ensures that it can be fulfilled. Self-contradictory statements are unsatisfiable. It makes no sense to subject an unsatisfiable theorem to empirical testing, since it is impossible to find instances that corroborate the theorem.

In the context of definitions $\Sigma_{\mathsf{def}} \subseteq \Sigma$, a theorem φ is *satisfiable* if there exists a model $\mathcal{A}$ in which the definitions hold, and the theorem is true: $\mathcal{A} \models \Sigma_{\mathsf{def}}$ and $\mathcal{A} \models \varphi$. A theorem φ is *unsatisfiable* if we can derive a contradiction from only the set of definitions and the theorem: $\Sigma_{\mathsf{def}} \cup \{\varphi\} \vdash \perp$.

Contingency Theorems that can both be fulfilled and refuted are called *contingent*—their validity strictly depends on the axioms, they are neither tautologically true, nor self-contradictory. The empirical investigation of non-contingent theorems does not make any sense because the outcome is predetermined.

A theorem is *contingent* if it is both satisfiable and falsifiable. A theorem is *non-contingent* if it is unsatisfiable or unfalsifiable (or both).

Further advantages Making the inference structure of a theory explicit will make it possible to assess the theory's *explanatory and predictive power* (by looking at the set of theorems because these are the predictions of the theory, and the proofs give explanations for them); its *domain* or scope (by investigating the models of the theory); its *coherence* (for example, a theory may turn out to have unrelated or independent parts); its *parsimony* (for example, it may turn out that some assumptions are not necessary, or can be relaxed); and other properties.

Computational tools

The above operationalizations require particular proof or model searches for establishing the criteria. The field of automated reasoning has provided us with automated theorem provers and model generators—generic tools that can directly be used for computational testing of the criteria. Automated theorem provers, such as OTTER (McCune 1994b),

are programs that are designed to find proofs of theorems. Typical theorem provers use *reductio ad absurdum*, that is, the program attempts to derive a contradiction from the premises and the negation of the theorem. A theorem prover can also be used to prove that a theory is inconsistent if it can derive a contradiction from the set of premises of the theory. Automated model generators, such as MACE (McCune 1994a), are programs that can find (small) models of sets of sentences. A model generator can prove the consistency of a theory, if it can generate a model of the premises. It can also be used to prove underivability of a conjecture, by attempting to generate a model of the premises in which the conjecture is false. Table 1 summarizes how to test for the criteria.[2]

CRITERION	OTTER	MACE
Consistency		$(\exists\mathcal{A})\mathcal{A} \models \Sigma$
Inconsistency	$\Sigma \vdash \bot$	
Soundness	$\Sigma \cup \{\neg\varphi\} \vdash \bot$	
Unsoundness		$(\exists\mathcal{A})\mathcal{A} \models \Sigma \cup \{\neg\varphi\}$
Falsifiability		$(\exists\mathcal{A})\mathcal{A} \models \Sigma_{\mathsf{def}} \cup \{\neg\varphi\}$
Unfalsifiability	$\Sigma_{\mathsf{def}} \cup \{\neg\varphi\} \vdash \bot$	
Satisfiability		$(\exists\mathcal{A})\mathcal{A} \models \Sigma_{\mathsf{def}} \cup \{\varphi\}$
Unsatisfiability	$\Sigma_{\mathsf{def}} \cup \{\varphi\} \vdash \bot$	

Note: Σ denotes a premise set, with Σ_{def} the definitions in this set, and φ a conjecture or theorem.

Table 1: Criteria and Automated Reasoning Tools.

The decision to use either automated theorem provers or model generators is not an arbitrary one. Although (syntactic) proof-theoretic and (semantic) model-theoretical characterizations are logically equivalent, determining the criteria is a different matter. First, there is a fundamental restriction on what we can hope to achieve because first order logic is undecidable. Although undecidable, first-order logic is semi-decidable—we can prove $\Sigma \vdash \bot$ if it is true. We suggest the use of theorem provers only for cases in which a contradiction can be derived. The other cases are, in general, undecidable (causing theorem provers to run on for ever). However, finding finite models is, again, decidable—we can find a finite model $\mathcal{A}$ such that $\mathcal{A} \models \Sigma$ if there exists such a finite model. Current model generators can only find finite models—even only models of small cardinalities. We have no solution for those cases in which only infinite models exist (or only models that are too large for current programs). Second, the tests we suggest to evaluate the criteria do not only prove a criterion, but also present a specific proof or model that is available for further inspection. In simple cases theorem provers may terminate after exhausting their search space without finding a contradiction, proving indirectly that the problem set is consistent, or that a conjecture is not derivable. Even in these cases a direct proof is far more informative, for example, if we can find specific counterexamples to a conjecture, it is immediately clear why our proof attempt has failed.

[2] OTTER and MACE are companion programs that can read the same input format. This facilitates switching between theorem proving and model searching, depending on which type of tool is most suitable for the specific proof or disproof attempt at hand. Kamps (1998) discusses the use of these tools for formal theory building in detail.

Case Study: A Formal Theory of Organizations in Action

We will illustrate the criteria outlined above by applying them to the formal theory of Organizations in Action (Kamps & Pólos 1999), a formal rendition of (Thompson 1967). Thompson (1967) is one of the classic contributions to organization theory: it provides a framework that unifies the perspective treating organizations as closed systems, with the perspective that focuses on the dependencies between organizations and their environment. This framework has influenced much of the subsequent research in organization theory. Thompson (1967) is a ordinary language text, in which only the main propositions are clearly outlined. Kamps and Pólos (1999) provide a formal rendition of the first chapters of the book, by reconstructing the argumentation used in the text.

PREDICATES	
Primitive:	
$\mathsf{O}(o)$	o is an organization
$\mathsf{SO}(o, so)$	so is a suborganization of o
$\mathsf{TC}(o, tc)$	tc is the technical core of o
$\mathsf{REVA}(o, tc)$	tc is rationally evaluated by o
$\mathsf{UC}(o, u)$	o has uncertainty u
$\mathsf{RED}(o, i, tc)$	o attempts to reduce i for tc
$\mathsf{FL}(tc, f, o)$	tc is exposed to a fluctuation f from o
$\mathsf{CS}(tc, c, o)$	tc is exposed to a constraint c from o
$\mathsf{C}(i, u)$	i causes u
$\mathsf{HC}(o, i)$	o has control over i
Defined:	
$\mathsf{CO}(o)$	o is a complex organization
$\mathsf{ENVI}(tc, i, o)$	tc is exposed to an influence i from o
$\mathsf{SEFF}(o, i, tc)$	o seals off tc from i
$\mathsf{ATO}(o)$	o is an atomic organization
$\mathsf{BUF}(o, f, tc)$	o buffers f for tc
$\mathsf{ANA}(o, c, tc)$	o anticipates and adapts to c for tc
$\mathsf{CEE}(o_1, o_2)$	o_2 is in o_1's controlled environment
$\mathsf{SM}(o, f, tc)$	o smoothes f for tc
$\mathsf{NEG}(o, c, tc)$	o negotiates c for tc

Table 2: Predicates (Kamps & Pólos 1999).

The formal theory uses the predicate symbols reprinted in table 2. Although the text of (Thompson 1967) does not contain explicit definitions, the use of terminology in the text strongly suggests strict dependencies between several important concepts. This allowed for the definition of these concepts in terms of a small number of primitive notions of organization theory. Table 3 lists the premises (both definitions and assumptions) and theorems of the formal theory. In the formal theory, the key propositions of (Thompson 1967) can be derived as theorems (theorem 3, corollaries 8, 9, and 11). The proofs of these theorems are based on a reconstruction of the argumentation in the text (assumptions 1–8). Additionally, the formal theory explains why the theory is restricted to a particular type of organizations (theorem 6). Moreover, it derives a heretofore unknown implication of the theory (corollary 12) that relates Thompson's theory to recent empirical findings and current developments in organization theory. For detailed discussion we refer the reader to (Kamps & Pólos 1999).

The criteria of the previous section played an important

PREMISES AND THEOREMS	
Premises:	
Def.1	$\forall x\ [\mathsf{CO}(x) \leftrightarrow \mathsf{O}(x) \wedge \exists y\ [\mathsf{SO}(x,y) \wedge \mathsf{TC}(x,y)]]$
Def.2	$\forall x,y,z\ [\mathsf{ENVI}(x,y,z) \leftrightarrow \mathsf{FL}(x,y,z) \vee \mathsf{CS}(x,y,z)]$
Def.3	$\forall x,y,z\ [\mathsf{SEFF}(x,y,z) \leftrightarrow \mathsf{SO}(x,z) \wedge \exists v,w\ [\mathsf{ENVI}(z,y,v) \wedge \mathsf{UC}(z,w) \wedge \mathsf{C}(y,w) \wedge \mathsf{RED}(x,w,z)]]$
Def.4	$\forall x\ [\mathsf{ATO}(x) \leftrightarrow \mathsf{O}(x) \wedge \neg\mathsf{CO}(x)]$
Def.5	$\forall x,y,z\ [\mathsf{BUF}(x,y,z) \leftrightarrow \mathsf{SO}(x,z) \wedge \mathsf{FL}(z,y,x) \wedge \mathsf{RED}(x,y,z)]$
Def.6	$\forall x,y,z\ [\mathsf{ANA}(x,y,z) \leftrightarrow \mathsf{SO}(x,z) \wedge \mathsf{CS}(z,y,x) \wedge \mathsf{RED}(x,y,z)]$
Def.7	$\forall x,y\ [\mathsf{CEE}(x,y) \leftrightarrow \mathsf{O}(x) \wedge \forall z\ [\mathsf{ENVI}(x,z,y) \rightarrow \mathsf{HC}(x,z)]]$
Def.8	$\forall x,y,z\ [\mathsf{SM}(x,y,z) \leftrightarrow \mathsf{SO}(x,z) \wedge \exists v\ [\mathsf{FL}(x,y,v) \wedge \mathsf{RED}(x,y,z)]]$
Def.9	$\forall x,y,z\ [\mathsf{NEG}(x,y,z) \leftrightarrow \mathsf{SO}(x,z) \wedge \exists v\ [\mathsf{CS}(x,y,v) \wedge \mathsf{RED}(x,y,z)]]$
Ass.1	$\forall x,y,z\ [\mathsf{TC}(x,y) \wedge \mathsf{TC}(x,z) \rightarrow y = z]$
Ass.2	$\forall x,y\ [\mathsf{TC}(x,y) \rightarrow \mathsf{REVA}(x,y)]$
Ass.3	$\forall x,y,z\ [\mathsf{SO}(x,y) \wedge \mathsf{REVA}(x,y) \wedge \mathsf{UC}(y,z) \rightarrow \mathsf{RED}(x,z,y)]$
Ass.4	$\forall x,y,z\ [\mathsf{ENVI}(x,y,z) \rightarrow \exists v\ [\mathsf{UC}(x,v) \wedge \mathsf{C}(y,v)]]$
Ass.5	$\forall x,y\ [\mathsf{O}(x) \wedge \mathsf{TC}(x,y) \wedge \neg\mathsf{SO}(x,y) \rightarrow x = y]$
Ass.6	$\forall x,y,z,v,w\ [\mathsf{ENVI}(x,y,z) \wedge \mathsf{UC}(x,v) \wedge \mathsf{C}(y,v) \wedge \mathsf{UC}(x,w) \wedge \mathsf{C}(y,w) \rightarrow v = w]$
Ass.7	$\forall x,y,z,v\ [\mathsf{O}(x) \wedge \mathsf{SO}(x,y) \wedge \mathsf{ENVI}(y,z,v) \rightarrow \mathsf{HC}(x,z)]$
Ass.8	$\forall x,y,z,v\ [\mathsf{RED}(x,y,z) \wedge \mathsf{C}(v,y) \wedge \mathsf{HC}(x,v) \rightarrow \mathsf{RED}(x,v,z)]$
Theorems:	
Lem.1	$\forall x,y\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \rightarrow \mathsf{SO}(x,y)]$
Lem.2	$\forall x,y,z\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{UC}(y,z) \rightarrow \mathsf{RED}(x,z,y)]$
Thm.3	$\forall x,y,z,v\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{ENVI}(y,z,v) \rightarrow \mathsf{SEFF}(x,z,y)]$
Lem.4	$\forall x\ [\mathsf{ATO}(x) \leftrightarrow \mathsf{O}(x) \wedge \neg\exists y\ [\mathsf{SO}(x,y) \wedge \mathsf{TC}(x,y)]]$
Lem.5	$\forall x,y\ [\mathsf{ATO}(x) \wedge \mathsf{TC}(x,y) \rightarrow x = y]$
Thm.6	$\forall x,y,z,v,w\ [\mathsf{ATO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{ENVI}(y,v,z) \wedge \mathsf{UC}(y,w) \wedge \mathsf{C}(v,w) \rightarrow \exists u\ [\mathsf{UC}(x,u) \wedge \mathsf{C}(v,u) \wedge w = u]]$
Thm.7	$\forall x,y,z,v\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{ENVI}(y,z,v) \rightarrow \mathsf{RED}(x,z,y)]$
Cor.8	$\forall x,y,z\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{FL}(y,z,x) \rightarrow \mathsf{BUF}(x,z,y)]$
Cor.9	$\forall x,y,z\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{CS}(y,z,x) \rightarrow \mathsf{ANA}(x,z,y)]$
Thm.10	$\forall x,y,z,v,w\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{ENVI}(y,z,x) \wedge \mathsf{CEE}(x,v) \wedge \mathsf{ENVI}(x,w,v) \wedge \mathsf{C}(w,z) \rightarrow \mathsf{RED}(x,w,y)]$
Cor.11	$\forall x,y,z,v,w\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{FL}(y,z,x) \wedge \mathsf{CEE}(x,v) \wedge \mathsf{FL}(x,w,v) \wedge \mathsf{C}(w,z) \rightarrow \mathsf{SM}(x,w,y)]$
Cor.12	$\forall x,y,z,v,w\ [\mathsf{CO}(x) \wedge \mathsf{TC}(x,y) \wedge \mathsf{CS}(y,z,x) \wedge \mathsf{CEE}(x,v) \wedge \mathsf{CS}(x,w,v) \wedge \mathsf{C}(w,z) \rightarrow \mathsf{NEG}(x,w,y)]$

Table 3: A Formal Theory of Organizations in Action (Kamps & Pólos 1999).

	Lem.1	**Lem.2**	**Thm.3**	**Lem.4**	**Lem.5**	**Thm.6**	**Thm.7**	**Cor.8**	**Cor.9**	**Thm.10**	**Cor.11**	**Cor.12**
Consistent	yes											
Sound	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Falsifiable	yes	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes
Satisfiable	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Contingent	yes	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes

Table 4: Evaluating the Theory.

role during the construction of the formal theory. Table 4 give an assessment of the final version of the theory in terms of the criteria.

Consistency Using an automated model generator it is easy to find models of the theory. MACE produced a model of cardinality 4 within a second (a model having universe $\{0,1,2,3\}$, reprinted in tables 5 and 6). This is a prototypical model of the theory corresponding to the claims of theorem 3, theorem 7, and corollary 8. It represents an organization that seals its core technologies off from environmental fluctuations, by the use of buffering (for example the stockpiling of materials and supplies). It is easy to verify that all premises (and theorems) hold in the model—the model proves that the theory is consistent.

Finding any arbitrary model of the theory is, formally speaking, sufficient to prove its consistency. We can find models on smaller cardinalities. For example, on cardinality 1 there exists a trivial model that assigns *false* to all predicates. In practice, we try to find more natural models of the theory. That is, we can examine the models and see if they conform to our mental models of the theory. This is an easy safeguard against hidden inconsistencies—theories that are only consistent because background knowledge has remained implicit.[3] We can look for prototypical models of the theory directly by adding premises that express appropriate initial conditions (typically existential statements). If we find a model of this enlarged set of premises, it is obviously

[3]Consider the following simple example:

$$\forall x\ [\mathsf{Dog}(x) \rightarrow \mathsf{Bark}(x)]$$
$$\mathsf{Rottweiler}(\mathsf{Johnny})$$
$$\neg\mathsf{Bark}(\mathsf{Johnny})$$

Can we find a model of this theory? Yes, although inspection of the models will reveal that in every model of the theory, Johnny the rottweiler is not a dog. These models are nonintended models because of the (implicit) background knowledge that rottweilers are a particular breed of dogs.

$$\forall x\ [\mathsf{Rottweiler}(x) \rightarrow \mathsf{Dog}(x)]$$

Adding this assumption to the theory will make it inconsistent, that is, we can then derive a contradiction from it.

O	
0	T
1	F
2	F
3	F

SO	0	1	2	3
0	F	T	F	F
1	F	F	F	F
2	F	F	F	F
3	F	F	F	F

TC	0	1	2	3
0	F	T	F	F
1	F	F	F	F
2	F	F	F	F
3	F	F	F	F

REVA	0	1	2	3
0	F	T	F	F
1	F	F	F	F
2	F	F	F	F
3	F	F	F	F

UC	0	1	2	3
0	F	F	F	F
1	F	F	F	T
2	F	F	F	F
3	F	F	F	F

$\mathsf{RED}(0,3,1)$ = T and $\mathsf{RED}(0,2,1)$ = T and $\mathsf{RED}(x,y,z)$ = F otherwise.

$\mathsf{FL}(1,2,0)$ = T and $\mathsf{FL}(x,y,z)$ = F otherwise.

$\mathsf{CS}(x,y,z)$ = F for all x, y, z in $\{0,1,2,3\}$.

C	0	1	2	3
0	F	F	F	F
1	F	F	F	F
2	F	F	F	T
3	F	F	F	F

HC	0	1	2	3
0	F	F	T	F
1	F	F	F	F
2	F	F	F	F
3	F	F	F	F

Table 5: A Model of the Theory (only primitives).

CO	
0	T
1	F
2	F
3	F

$\mathsf{ENVI}(1,2,0)$ = T and $\mathsf{ENVI}(x,y,z)$ = F otherwise.

$\mathsf{SEFF}(0,2,1)$ = T and $\mathsf{SEFF}(x,y,z)$ = F otherwise.

$\mathsf{BUFF}(0,2,1)$ = T and $\mathsf{BUFF}(x,y,z)$ = F otherwise.

Table 6: Selected Defined Predicates (extending Tab.5).

also a model of the theory.

Soundness In the final formal theory (as reprinted in table 3), all theorems are derivable. Figure 1 shows the inference structure of the formal theory. None of the proofs is very complex (the automated theorem prover OTTER required only 12 seconds for the longest proof).

The original text of the theory presupposes common background knowledge—assumptions taken for granted in the substantive field. In order for the theorems to be deductively derivable, several implicit assumptions had to be added to the theory (notably assumptions 1, 6 and 7). Which precise background assumptions to add is one of the thorniest problems in the formalization of a theory, requiring a deep understanding of the substantive field under consideration. Fortunately, the formal tools can help: suppose we cannot derive a theorem due to a missing background assumption. We can prove that the theorem is underivable by generating counterexamples, that is, models of the premises in which the theorem is false. If the unsoundness of the theorem is due to missing background knowledge, inspection of the counterexamples will reveal that they are nonintended models—models that violate our common sense, or implicit background assumptions from the substantive domain. For example, we found models of organizations having more than one operational core (conflicting with assumption 1, which is implicit in the original text). We can make the theorem derivable if we add sufficient assumptions to exclude these nonintended models from the theory.

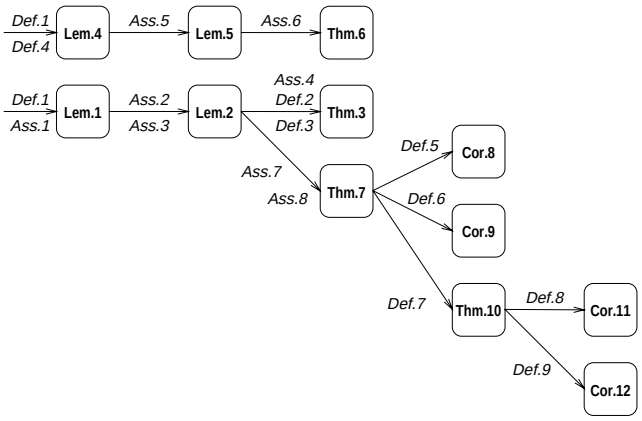

Figure 1: Inferences in the Formal Theory

Falsifiability We tried to prove the falsifiability of the theorems as discussed above: by finding a model in which the definitions hold and the theorem is false. We failed to find such a model for lemma 4. As it turns out, lemma 4 can be derived from just definitions 1 and 4—proving that this lemma is unfalsifiable. Lemma 4 is true by definition, and therefore does not make an empirical claim. If we would subject lemma 4 to empirical testing, we will be unable to refute it, but can at best reassert the trivial validity of the statement.

Fortunately, the other theorems of the formal theory are falsifiable. For each of these theorems, we can find models of the definitions in which the theorem is false (not reproduced here). MACE generated these models in a matter of seconds.

Satisfiability/Contingency For proving the satisfiability of the theorems, we need to find models that make both the definitions and the theorem true. The model of tables 5 and 6 also proves the satisfiability of all theorems. As a result, we can conclude that only lemma 4 is non-contingent—it is not an empirical statement, but its truth is determined by virtue of the definitions only. The other theorems make empirical claims that can, in principle, be corroborated or refuted by empirical testing.

Conclusions and Discussion

This paper discussed the axiomatization of scientific theories in first order logic. We provided practical operationalizations of criteria for evaluating scientific theories, such as the consistency and falsifiability of theories and the soundness of inferences. The precise formulation of these criteria is tailored to the use of computational support. The tests for the criteria, in practice amounting to particular proof or model searches, can be directly performed by existing automated reasoning tools.

The efficient treatment of definitions is one of the basic research problems in automated reasoning (Wos 1988, Problem 30). A naive approach is to eliminate all defined

predicates and functions from the problem set by expanding the definitions. This, however, also eliminates useful ways chunking information and as a result it "increases the likelihood that a program will get lost" (Wos 1988, p.62).[4] Since only few problems are provable without expanding (some of) the definitions, dealing with definitions is a difficult challenge for automated reasoning tools. As a result, most automated theorem provers treat definitions and axioms alike (a notable exception is (Giunchiglia & Walsh 1989)). Interestingly, the above argument does not seem to apply to automated model generators.[5] The search space of automated model generators is the set of all possible models, i.e., all possible interpretation functions of the logical vocabulary. Reducing the vocabulary of the formal language by eliminating the defined concepts will proportionally reduce this search space. Moreover, after eliminating the defined concepts, the definitions themselves can be removed from the problem set, which reduces the number of "constraints" that need to be taken into account when deciding whether a particular interpretation is a model of the problem set.

We used the criteria to evaluate a formal rendition of a classic organization theory (Kamps & Pólos 1999). Assessing the criteria allows for an exact evaluation of the merits of a theory. In some cases, this may reveal important facts about the theory, for example, the case study showed that one of the derived statements is unfalsifiable—empirical investigation of it is futile. However, we do not view the criteria as rigid, final tests. Quite the opposite, in our experience the criteria are especially useful during the process of formalizing a theory. During the construction of a formal theory, the criteria can provide useful feedback on how to revise the theory in case of a deficiency. For example, examining counterexamples can reveal which implicit (background) assumptions need to be added to the theory. There are, of course, important principled and practical limitations to the axiomatization of theories in first order logic: the undecidability of first order logic, the scientific knowledge available in the substantive domains, or the availability of resources like processor power, memory, and time. There is yet no equivocal answer to the question whether it is possible, or even desirable, to axiomatize large parts of substantive domains. Axiomatization is often viewed as the ultimate step in the lifetime of a scientific theory—the axioms are frozen in their final form, and active research moves on to areas where still progress can be made. The main motivation for the research reported in this paper is that the formalization of theories can play a broader role: it need not end the life of a theory, but rather contribute to its further development.

Acknowledgments Thanks to Michael Masuch, László Pólos, Ivar Vermeulen, and the *AAAI* reviewers for discussion and insightful comments, and to Mary Beth Jenssen without whom this paper would not have been possible.

[4]This is substantiated by the theory in our case study: after eliminating all defined predicates, OTTER proved several theorems slightly faster, but some others slower. These are preliminary observations without taking into account the time needed expand the definitions (a preprocessing step that is done once for any number of queries). This expansion was done manually but is of no great complexity: since definitions are not allowed to be circular, we only need to expand each definition once.

[5]For the theory in our case study, eliminating defined concepts gave significantly better performance on all model searches (roughly halving MACE's processor and memory usage). Again, these are preliminary observations that do not consider the preprocessing of definitions.

References

Bruggeman, J. 1997. Niche width theory reappraised. *Journal of Mathematical Sociology* 22(2):201–220.

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

Fisher, A. 1988. *The logic of real arguments*. Cambridge University Press, Cambridge UK.

Giunchiglia, F., and Walsh, T. 1989. Theorem proving with definitions. In Cohn, A., and Dennison, J., eds., *AISB-89: Proceedings of the seventh conference of Artificial Intelligence and the Simulation of Behaviour*, 175–183. Pitman, London.

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

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

Kamps, J., and Pólos, L. 1999. Reducing uncertainty: A formal theory of *Organizations in Action*. *American Journal of Sociology* 104(6):1774–1810.

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

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

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

Péli, G., and Masuch, M. 1997. The logic of propagation strategies: Axiomatizing a fragment of organizational ecology in first-order logic. *Organization Science* 8:310–331.

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

Péli, G. 1997. The niche hiker's guide to population ecology: A logical reconstruction of organizational ecology's niche theory. In Raftery, A. E., ed., *Sociological Methodology 1997*, 1–46. Blackwell, Oxford UK.

Popper, K. R. 1959. *The Logic of Scientific Discovery*. Hutchinson, London.

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

Thompson, J. D. 1967. *Organizations in Action: Social Science Bases of Administrative Theory*. McGraw-Hill, New York.

Wos, L. 1988. *Automated Reasoning, 33 Basic Research Problems*. Prentice Hall, Englewood Cliffs, New Jersey.