COMPARING THEORIES VIA INTERPRETATION

RICHARD KIMBERLY HECK

1. INTERPRETATION

One of the central concerns of more advanced logic is relations of relative strength between theories. There’s an obvious sense in which, say, PA is stronger than Q, namely, that PA proves all of the Q’s theorems and more. In that sort of case, we say that PA contains or extends Q. But this method of comparison only works when the two theories have the same language. Otherwise, the two theories will always prove disjoint sets of theorems, if only because they will prove different instances of the law of excluded middle: If A is a sentence of the language of some theory Σ₁ but not of Σ₂, then Σ₁ will prove A ∨ ¬A, but Σ₂ will not. If, however, we have a language L₁ that contains (extends) L₂, then we have another option: If Σ₁ is a theory in L₁ and Σ₂ is a theory in L₂, we can ask whether Σ₁ proves anything that Σ₂ might but doesn’t prove, that is, whether there is some sentence A of L₂ such that Σ₁ ⊢ A but Σ₂ ⊬ A. If there is not, then we say that Σ₁ is a conservative extension of Σ₂.

But, as said, this only works in the case where the language of Σ₁ contains that of Σ₂. It would not allow us, for example, to compare the strengths of theories stated in disjoint languages. But we might well want to be able to do that. What is now the established method for doing so was first studied in detail by Tarski around 1950, but the idea are implicit in much work in mathematics in the late 1800s. (As we shall see, this method can, in fact, be applied even when the theories we are comparing are stated in the same language, so it is actually a fully general way to compare the strength of theories.) The basic idea is that B (for ‘base’) is at least as strong as T (for ‘target’) if it is possible to define the primitives of T in terms of those of B in such a way that B is then able to prove the translations of T’s theorems. For example, Zermelo-Frankel set theory, ZF, is at least as strong as Peano Arithmetic, PA because there is a way of defining the 0, S, +, ×, and < in terms of ∈ such that ZF is then able to prove the translations of all of PA’s theorems. But PA is not as strong as ZF because there is no way of defining ∈ in terms of 0, S, +, ×, and < such that PA is then able to prove the translations of all the theorems of ZF.¹

We need to talk first, then, about what we mean by a ‘possible definition’. The simplest case is that of relation symbols. If R(x₁,...,xₙ) is an n-place relation symbol from L(T), then a ‘possible definition’ of it has the form:

\[ R(x₁,...,xₙ) \equiv \phi_R(x₁,...,xₙ) \]

where \( \phi_R(x₁,...,xₙ) \) is some formula of \( L(B) \) containing exactly the mentioned variables free.² The case of function symbols is more complicated, for the familiar

¹As it happens, we can do this for weaker theories. In particular, it can be shown that PA interprets a theory known as hereditarily finite set theory, which is ZF without the axiom of infinity. The converse is also true, so these two theories are, as it is said, ‘mutually interpretable’.

²A more general notion allows for parameters, but we shall not consider that here.
reason that there simply may not be the right sorts of terms for us to give a
definition of the form:

\[ f(x_1, \ldots, x_n) = t_f(x_1, \ldots, x_n) \]

where \( t(x_1, \ldots, x_n) \) is some term of \( \mathcal{L}(B) \). There is no hope, for example, of
producing a term of the language of set theory to define \( + \): The only terms in the
language of set theory are the variables. What we can do instead, though, is to
define the graph of the successor function (much as we ‘represent’ functions by
representing their graphs). A ‘possible definition’ of a function symbol will thus
take the following form:

\[ f(x_1, \ldots, x_n) = y \equiv \phi_f(x_1, \ldots, x_n, y) \]

But of course we need \( \phi_f \) to define a function, so we require that \( B \) prove existence
and uniqueness:

\[
B \vdash \forall x_1 \ldots \forall x_n \exists y[\phi_f(x_1, \ldots, x_n, y)]
\]

\[
B \vdash \forall x_1 \ldots \forall x_n \forall y \forall z[\phi_f(x_1, \ldots, x_n, y) \land \phi_f(x_1, \ldots, x_n, z) \rightarrow y = z]
\]

Constants may be treated as zero-place function symbols, so a definition of a
constant \( c \) takes the form:

\[
\forall x[c = x \equiv \phi_c(x)]
\]

In this case, we require \( B \) to be able to prove that \( \phi_c \) is uniquely satisfied:

\[
B \vdash \exists x(\phi_c(x) \land \forall y(\phi_c(y) \rightarrow x = y))
\]

Note that these ‘definitions’ are neither in \( \mathcal{L}(T) \) nor in \( \mathcal{L}(B) \) but rather in the
combined language \( \mathcal{L}(T) \cup \mathcal{L}(B) \).\(^3\)

A set of possible definitions for \( \mathcal{L}(T) \) in \( \mathcal{L}(B) \) is thus a set containing exactly
one formula of the mentioned form for each primitive (non-logical) symbol of
\( \mathcal{L}(T) \).

**Definition** (Tarski). \( B \) interprets \( T \) via \( D \) if:

(i) \( D \) is a set of possible definitions for \( \mathcal{L}(T) \) in \( \mathcal{L}(B) \);

(ii) \( B \) proves the existence and uniqueness claims associated with the definitions
of function and relation symbols;

(iii) \( B \cup D \) proves all of \( T \)'s theorems.

\( T \) is interpretable in \( B \) if \( B \) interprets \( T \) via \( D \), for some set of possible definitions
\( D \).

The definition just stated is the one Tarski originally gave, but we think of
things slightly differently nowadays. Any such a set of possible definitions will
allow us to translate formulas from \( \mathcal{L}(T) \) to \( \mathcal{L}(B) \) by applying those definitions recursively. We start by saying how to translate atomic formulas. This would be
easy if the only terms were variables: Then we could just replace \( R(v_1, \ldots, v_n) \)
with \( \phi_R(v_1, \ldots, v_n) \). If there are function symbols (or constants) present, however,
things become more complicated. Consider first a simple case:

\[
S(x) < y
\]

\(^3\)If the languages overlap—if, say, \( S \) is a function symbol in both languages—then we can just modify one of the theories so that the function symbol is not \( S \) but, say, \( S^* \).
First, we put this into a form in which the function symbol $S$ appears only in formulae of the form $S(x) = z$, thus:

\[(1.2) \quad \exists z (f(x) = z \land z < y)\]

Note that this is logically equivalent to (1.1). We can then replace $f(x) = y$ by what defines it, thus:

\[(1.3) \quad \exists z (\phi_S(x, z) \land z < y)\]

This is now not logically equivalent to (1.2), but it does follow from the ‘possible definition’ $f(x) = z \equiv \phi_S(x, z)$ that (1.3) is equivalent to (1.2). We can then replace $z < y$ with what defines it:

\[(1.4) \quad \exists z (\phi_S(x, z) \land \phi_<(z, y))\]

So this says, roughly: There is a thing that is the value of $S$ for argument $x$, and that thing is less than $y$. Note again that the ‘possible definitions’ imply that (1.4) is equivalent to (1.1).

A more complicated case would be: $S(x + y) < Sy$. We start by working on the left-most term and work from outside in and, again, left to right. So:

\[S(x + y) < y + y\]

\[\exists z \{S(x + y) = z \land z < y + y\}\]

\[\exists z \{\exists w [w = x + y \land S(w) = z \land z < y + y]\}\]

\[\exists z \exists w \exists v \{w = x + y \land S(w) = z \land v = y + y \land z < v\}\]

The last line here is again logically equivalent to the first one. We can now replace all the identities involving function symbols with what defines them:

\[\exists z \exists w \exists v \{\phi_+(x, y, w) \land \phi_S(w, z) \land \phi_+(y, y, v) \land z < v\}\]

Finally, we now can replace $z < u$ with what defines it:

\[\exists z \exists w \exists v \{\phi_+(x, y, w) \land \phi_S(w, z) \land \phi_+(y, y, v) \land \phi_<(z, v)\}\]

And, once again, though this is no longer logically equivalent to what preceded it, the equivalence does follow from the ‘possible definitions’ we employed.

That, then, explains how to translate atomic formulas in accord with some set of ‘possible definitions’.

We obviously could have done this in a number of different ways. In a sense, it really doesn’t matter exactly how we do it. Since B proves existence and uniqueness for all the formulae we are using to translate function symbols and constants, all reasonable translations will be B-provably equivalent. We will want to be able to talk, though, about the translation of a given formula, so we will suppose that we have specified some particular order in which the translation is to proceed—left to right, outside in, say—and that we have also stated, in advance, what choices we are to make at any point that such a choice might need making (e.g., what variables are to be used at each stage of the process).
Let $A^*$ be the translation of $A$. Then we extend the translation to all formulas as follows: $^4$

$$(\neg A)^* = \neg (A^*)$$

$$(A \to B)^* = (A^*) \to (B^*)$$

$$[\forall v(A(v))]^* = \forall v(A(v)^*)$$

Similarly for $\lor$, $\land$, and $\exists$. The crucial point about these translations are that they preserves logical structure. If a formula is, say, of the form $\forall x[\exists y A(x, y) \to B(x)]$, then its translation will also be of that same form—even though the atomic formulae of which the original formula is composed may well have been replaced by formulae that are not atomic.

It will be important below that the translation function is recursive: Given a formula, one can calculate its translation simply by following the procedure just described. $^5$ (It can, moreover, be shown that the translation function is one-one and that its inverse is recursive, though that will not be essential to what follows.)

The more modern way to think of interpretation, then, is simply in terms of this notion of translation. There is still a place for ‘possible definitions’, but do not think of the definitions as constituting a theory linking $T$ with $B$ but only as the foundation on which to build a translation from $L(T)$ to $L(B)$. So we can redefine the notion of interpretation as follows.

**Definition 1.1.** $B$ interprets $T$ via $D$ if

(i) $D$ is a set of possible definitions of the primitives of $L(T)$ in terms of those of $L(B)$; $^6$

(ii) $B$ proves the existence and uniqueness claims associated with the definitions of function and relation symbols;

(iii) For each axiom $A$ of $T$, $B \vdash A^*$, where $(\cdot)^*$ is the translation from $L(T)$ to $L(B)$ that $D$ induces.

Note that (iii) is different from what was stated above: The corresponding clause here would have said that for each theorem $A$ of $T$, $B \vdash A^*$. But the weaker clause implies the stronger, as the following shows.

**Theorem 1.2.** Suppose If $B$ interprets $T$ via $D$, then, if $T \vdash A$, then $B \vdash A^*$; i.e., $B$ proves the translations of all $T$’s theorems.

Note that we do not necessarily have that $B \vdash A^*$ only if $T \vdash A$. If that does hold, then we say that the interpretation is faithful (though this notion will play no role in what follows).

The proof of Theorem 1.2 is fussy but not difficult. Given a $T$-proof of $A$, we first consider the result of simply translating the proof step-by-step into $L(B)$. That will not itself be a $B$-proof, but it can be transformed into one. There will typically be steps in the original $T$-proof where we have appealed to some axiom $A$ of $T$. At the corresponding point in the translated proof, we will have the translation of that axiom, $A^*$. This is not necessarily an axiom of $B$, but it is, by hypothesis, a theorem of $B$. So we can replace the single step in the original

---

$^4$That is: The translation of the negation of $A$ is the negation of the translation of $A$, etc.

$^5$If $L(T)$ is infinite, then we need to require that the set of possible definitions is also recursive.

$^6$Actually, ‘possible definitions’ now are not biconditionals, but just translations. But we shall not pursue the point.
proof with a $B$-proof of $A^*$. The case of logical axioms and rules of inference is essentially the same as in the earlier proof: Assuming that $\mathcal{L}(T)$ contains neither constants nor function symbols, the translations of instances of logical axioms will still be instances of logical axioms, and the rules of inference will still apply to the translations of the formulas to which they applied in the original proof. If $\mathcal{L}(T)$ does contain constants or function symbols, then matters will be a bit more complicated, but it can be shown that the proof can again be patched so as to work.

We thus have the following corollary.

**Corollary 1.3.** Suppose that $T$ is interpretable in $B$. Then, if $B$ is consistent, then so is $T$.

**Proof.** Suppose that $T$ is inconsistent. Then for some formula $A$ of $\mathcal{L}(T)$, $T \vdash A \land \neg A$. So, by Theorem 1.2, $B \vdash (A \land \neg A)^*$. But $(A \land \neg A)^* = A^* \land \neg (A^*)$, so $B \vdash A^* \land \neg (A^*)$, and $B$ is inconsistent. \hfill $\Box$

2. R**E**LATIVE I**NTERPRETATION

The notion of interpretation, in the sense in which we have been discussing it recently, is syntactic: To show that one theory $T$ is interpretable in another theory $B$, one describes a translation between the languages of the two theories—a syntactic transformation—and then shows that $B$ proves the translations of $T$’s axioms—a syntactic matter. There is, of course, also the semantic notion of interpretation. What is the relation between these two notions?

Suppose that $T$ is interpretable in $B$ via $D$, and let $\mathcal{M}$ be a model of $T$ and $\mathcal{N}$ a model of $B$. Then $D$ shows us, in effect, how to build a model of $T$ inside the model for $D$. The case of relation symbols is again the easiest. Suppose $R(x_1, \ldots, x_n)$ is an $n$-place relation symbol from $\mathcal{L}(T)$. Then there is some formula from $\mathcal{L}(B)$ $\phi_R(x_1, \ldots, x_n)$ that defines it. We may simply take the extension of $R$, in the new model, to be that of $\phi_R(x_1, \ldots, x_n)$. Similarly for function symbols and constants.

And since $B$ proves the translations of all $T$’s axioms, this really is a model of $T$.

There is one element of the semantic notion of interpretation, though, that is not reflected in the syntactic notion, namely, the idea of a domain. And we will need this in certain cases. Suppose, for example, that we want to interpret $\text{PA}$ in $\text{ZF}$. Then we might define $0$ as the empty set and let the successor of a number be the singleton of that number: \footnote{Of course, we cannot say $x = \emptyset$ in the language of set theory, since $\emptyset$ is not a symbol in that language.}

\[
\begin{align*}
x &= 0 \equiv \neg \exists y (y \in x) \\
x &= Sy \equiv x \in y \land \forall z (z \in y \rightarrow z = x)
\end{align*}
\]

Then we will want the translations of $\text{PA}$’s axioms somehow to reflect the fact that the ‘numbers’, as we are understanding them, are just $\emptyset$, $\{\emptyset\}$, $\{\{\emptyset\}\}$, and so on. In particular, not every set is a number.

In fact, however, it is fairly easy to allow for this sort of thing. We may take the domain to be given by some formula $\delta(x)$ of $\mathcal{L}(B)$, and we then translate the quantifiers by relativizing them:

\[
\begin{align*}
[\forall v(A(v))]^* &= \forall v(\delta(v) \rightarrow A(v)^*) \\
[\exists v(A(v))]^* &= \exists v(\delta(v) \land A(v)^*)
\end{align*}
\]

\begin{align*}
\end{align*}
So the translation of ∀v(A(v)) says, in effect, that every object in the ‘domain’ given by δ(x) satisfies the translation of A(v).

This gives rise to the notion of relative interpretation.

**Definition 2.1.** B relatively interprets T via D and δ(x) if

(i) D is a set of possible definitions of the primitives of L(T) in terms of those of L(B);

(ii) B proves the existence and uniqueness claims associated with the definitions of function and relation symbols, and also proves ∃x(δ(x));\(^8\)

(iii) For each axiom A of T, B ⊢ A\(^∗\), where (·)\(^∗\) is the translation from L(T) to L(B) that D induces.

**Theorem 2.2.** Suppose If B relatively interprets T via D, then, if T ⊢ A, then B ⊢ A\(^∗\); i.e., B proves the translations of all T’s theorems.

The proof is essentially the same as that of Theorem 1.2, with additional complications due to the relativization of the quantifiers.

**Corollary 2.3.** Suppose that T is relatively interpretable in B. Then, if B is consistent, then so is T.

\(^8\)The latter, of course, corresponds to the usual requirement that the domain be non-empty.