1 Invitation to Metatheory

This chapter is meant to serve as a preview, and for motivation to work through the chapters to come. In the next chapter, we'll move quickly into "categorical set theory" – which isn't all that difficult, but which is not yet well known among philosophers. For the past fifty years or so, it has almost been mandatory for analytic philosophers to know a little bit of set theory. However, it has most certainly not been mandatory for philosophers to know a little bit of category theory. Indeed, most analytic philosophers are familiar with the words "subset" and "powerset" but not the words "natural transformation" or "equivalence of categories." Why should philosophers bother learning these unfamiliar concepts?

The short answer is that is that category theory (unlike set theory) was designed to explicate *relations* between mathematical structures. Since philosophers want to think about relations between theories (e.g., equivalence, reducibility) and since theories can be modeled as mathematical objects, philosophers' aims will be facilitated by gaining some fluency in the language of category theory. At least that's one of the main premises of this book. So, in this chapter, we'll review some of the basics of the metatheory of propositional logic. We will approach the issues from a slightly different angle than usual, placing less emphasis on what individual theories say and more emphasis on the relations between these theories.

To repeat, the aim of **metatheory** is to theorize about theories. For simplicity, let's use M to denote this hypothetical theory about theories. Thus, M is not the *object* of our study; it is the *tool* we will use to study other theories and the relations between them. In this chapter, I will begin using this tool M to talk about theories – without explicitly telling you anything about M itself. In the next chapter, I'll give you the user's manual for M.

1.1 Logical Grammar

DEFINITION 1.1.1 A **propositional signature** Σ is a collection of items, which we call **propositional constants**. Sometimes these propositional constants are also called **elementary sentences**. (Sometimes people call them atomic sentences, but we will be using the word "atomic" for a different concept.)

These propositional constants are assumed to have no independent meaning. Nonetheless, we assume a primitive notion of identity between propositional constants; the fact

https://doi.org/10.1017/9781316275603.003 Published online by Cambridge University Press

that two propositional constants are equal or non-equal is not explained by any more fundamental fact. This assumption is tantamount to saying that Σ is a **bare set** (and it stands in gross violation of Leibniz's principle of the identity of indiscernibles).

ASSUMPTION 1.1.2 The **logical vocabulary** consists of the symbols $\neg, \wedge, \vee, \rightarrow$. We also use two further symbols for punctuation: a left and a right parenthesis.

DEFINITION 1.1.3 Given a propositional signature Σ, we define the set $\mathsf{Sent}(\Sigma)$ of Σ-sentences as follows:

1. If $\phi \in \Sigma$, then $\phi \in \mathsf{Sent}(\Sigma)$.
2. If $\phi \in \mathsf{Sent}(\Sigma)$, then $(\neg\phi) \in \mathsf{Sent}(\Sigma)$.
3. If $\phi \in \mathsf{Sent}(\Sigma)$ and $\psi \in \mathsf{Sent}(\Sigma)$, then $(\phi \wedge \psi) \in \mathsf{Sent}(\Sigma)$, $(\phi \vee \psi) \in \mathsf{Sent}(\Sigma)$, and $(\phi \rightarrow \psi) \in \mathsf{Sent}(\Sigma)$.
4. Nothing is in $\mathsf{Sent}(\Sigma)$ unless it enters via one of the previous clauses.

The symbol ϕ here is a variable that ranges over finite strings of symbols drawn from the alphabet that includes Σ; the connectives $\neg, \wedge, \vee, \rightarrow$; and (when necessary) left and right parentheses "(" and ")". We will subsequently play it fast and loose with parentheses, omitting them when no confusion can result. In particular, we take a negation symbol $\neg$ always to have binding precedence over the binary connectives.

Note that each sentence is, by definition, a finite string of symbols and, hence, contains finitely many propositional constants.

Since the set $\mathsf{Sent}(\Sigma)$ is defined inductively, we can prove things about it using "proof by induction." A proof by induction proceeds as follows:

1. Show that the property of interest, say P, holds of the elements of Σ.
2. Show that if P holds of ϕ, then P holds of $\neg\phi$.
3. Show that if P holds of ϕ and ψ, then P holds of $\phi \wedge \psi$, $\phi \vee \psi$, and $\phi \rightarrow \psi$.

When these three steps are complete, one may conclude that all things in $\mathsf{Sent}(\Sigma)$ have property P.

DEFINITION 1.1.4 A **context** is essentially a finite collection of sentences. However, we write contexts as sequences, for example $\phi_1, \ldots, \phi_n$ is a context. But ϕ_1, ϕ_2 is the same context as ϕ_2, ϕ_1, and is the same context as ϕ_1, ϕ_1, ϕ_2. If Δ and Γ are contexts, then we let Δ, Γ denote the union of the two contexts. We also allow an empty context.

1.2 Proof Theory

We now define the relation $\Delta \vdash \phi$ of derivability that holds between contexts and sentences. This relation is defined **recursively** (aka **inductively**), with base case $\phi \vdash \phi$ (Rule of Assumptions). Here we use a horizontal line to indicate that if $\vdash$ holds between the things above the line, then $\vdash$ also holds for the things below the line.

https://doi.org/10.1017/9781316275603.003 Published online by Cambridge University Press

Rule of Assumptions

$$\frac{}{\phi \vdash \phi}$$

$\wedge$ elimination

$$\frac{\Gamma \vdash \phi \wedge \psi}{\Gamma \vdash \phi} \qquad \frac{\Gamma \vdash \phi \wedge \psi}{\Gamma \vdash \psi}$$

$\wedge$ introduction

$$\frac{\Gamma \vdash \phi \qquad \Delta \vdash \psi}{\Gamma, \Delta \vdash \phi \wedge \psi}$$

$\vee$ introduction

$$\frac{\Gamma \vdash \phi}{\Gamma \vdash \phi \vee \psi} \qquad \frac{\Gamma \vdash \psi}{\Gamma \vdash \phi \vee \psi}$$

$\vee$ elimination

$$\frac{\Gamma \vdash \phi \vee \psi \qquad \Delta, \phi \vdash \chi \qquad \Theta, \psi \vdash \chi}{\Gamma, \Delta, \Theta \vdash \chi}$$

$\rightarrow$ elimination

$$\frac{\Gamma \vdash \phi \rightarrow \psi \qquad \Delta \vdash \phi}{\Gamma, \Delta \vdash \psi}$$

$\rightarrow$ introduction

$$\frac{\Gamma, \phi \vdash \psi}{\Gamma \vdash \phi \rightarrow \psi}$$

RA

$$\frac{\Gamma, \phi \vdash \psi \wedge \neg\psi}{\Gamma \vdash \neg\phi}$$

DN

$$\frac{\Gamma \vdash \neg\neg\phi}{\Gamma \vdash \phi}$$

The definition of the turnstyle $\vdash$ is then completed by saying that $\vdash$ is the smallest relation (between sets of sentences and sentences) such that

1. $\vdash$ is closed under the previously given clauses, and
2. If $\Delta \vdash \phi$ and $\Delta \subseteq \Delta'$, then $\Delta' \vdash \phi$.

The second property here is called **monotonicity**.

There are a variety of ways that one can explicitly generate pairs Δ, ϕ such that $\Delta \vdash \phi$. A method for doing such is typically called a **proof system**. We will not explicitly introduce any proof system here, but we will adopt the following definitions.

DEFINITION 1.2.1 A pair Δ, ϕ is called a **sequent** or **proof** just in case $\Delta \vdash \phi$. A sentence ϕ is said to be **provable** just in case $\vdash \phi$. Here $\vdash \phi$ is shorthand for $\emptyset \vdash \phi$. We use $\top$ as shorthand for a sentence that is provable – for example, $p \rightarrow p$. We could then add as an inference rule "$\top$ introduction," which allowed us to write $\Delta \vdash \top$. It can be proven that the resulting definition of $\vdash$ would be the same as the original definition. We also sometimes use the symbol $\perp$ as shorthand for $\neg\top$. It might then be convenient to restate RA as a rule that allows us to infer $\Delta \vdash \neg\phi$ from $\Delta, \phi \vdash \perp$. Again, the resulting definition of $\vdash$ would be the same as the original.

DISCUSSION 1.2.2 The rules we have given for $\vdash$ are sometimes called the **classical propositional calculus** or just the **propositional calculus**. Calling it a "calculus" is

https://doi.org/10.1017/9781316275603.003 Published online by Cambridge University Press

meant to indicate that the rules are purely formal and don't require any understanding of the meaning of the symbols. If one deleted the DN rule and replaced it with Ex Falso Quodlibet, the resulting system would be the **intuitionistic propositional calculus**. However, we will not pursue that direction here.

1.3 Semantics

DEFINITION 1.3.1 An **interpretation** (sometimes called a **valuation**) of Σ is a function from Σ to the set {true, false}, i.e., an assignment of truth-values to propositional constants. We will usually use 1 as shorthand for "true" and 0 as shorthand for "false."

Clearly, an interpretation v of Σ extends naturally to a function $v : \mathsf{Sent}(\Sigma) \to \{0,1\}$ by the following clauses:

1. $v(\neg\phi) = 1$ if and only if $v(\phi) = 0$.
2. $v(\phi \wedge \psi) = 1$ if and only if $v(\phi) = 1$ and $v(\psi) = 1$.
3. $v(\phi \vee \psi) = 1$ if and only if either $v(\phi) = 1$ or $v(\psi) = 1$.
4. $v(\phi \to \psi) = v(\neg\phi \vee \psi)$.

DISCUSSION 1.3.2 The word "interpretation" is highly suggestive, but it might lead to confusion. It is sometimes suggested that elements of $\mathsf{Sent}(\Sigma)$ are part of an uninterpreted calculus without intrinsic meaning, and that an intepretation $v : \Sigma \to \{0,1\}$ endows these symbols with meaning. However, to be clear, $\mathsf{Sent}(\Sigma)$ and $\{0,1\}$ are both mathematical objects; neither one of them is more linguistic than the other, and neither one of them is more "concrete" than the other.

This point becomes even more subtle in predicate logic, where we might be tempted to think that we can interpret the quantifiers so that they range over all actually existing things. To the contrary, the domain of a predicate logic interpretation must be a *set*, and a set is something whose existence can be demonstrated by ZF set theory. Since the existence of the world is not a consequence of ZF set theory, it follows that the world is not a set. (Put slightly differently: a set is an abstract object, and the world is a concrete object. Therefore, the world is not a set.)

DEFINITION 1.3.3 A **propositional theory** T consists of a signature Σ, and a set Δ of sentences in Σ. Sometimes we will simply write T in place of Δ, although it must be understood that the identity of a theory also depends on its signature. For example, the theory consisting of a single sentence p is different depending on whether it's formulated in the signature $\Sigma = \{p\}$ or in the signature $\Sigma' = \{p,q\}$.

DEFINITION 1.3.4 (Tarski truth) Given an interpretation v of Σ and a sentence ϕ of Σ, we say that ϕ is **true** in v just in case $v(\phi) = 1$.

DEFINITION 1.3.5 For a set Δ of Σ sentences, we say that v is a **model** of Δ just in case $v(\phi) = 1$, for all ϕ in Δ. We say that Δ is **consistent** if Δ has at least one model, and that Δ is **inconsistent** if it has no models.

Any time we define a concept for sets of sentences (e.g., consistency), we can also extend that concept to theories, as long as it's understood that a theory is technically a pair consisting of a signature and a set of sentences in that signature.

DISCUSSION 1.3.6 The use of the word "model" here has its origin in consistency proofs for non-Euclidean geometries. In that case, one shows that certain non-Euclidean geometries can be translated into models of Euclidean geometry. Thus, if Euclidean geometry is consistent, then non-Euclidean geometry is also consistent. This kind of maneuver is what we now call a **proof of relative consistency**.

In our case, it may not be immediately clear what sits on the "other side" of an interpretation, because it's certainly not Euclidean geometry. What kind of mathematical thing are we interpreting our logical symbols into? The answer here – as will become apparent in Chapter 3 – is either a Boolean algebra or a fragment of the universe of sets.

DEFINITION 1.3.7 Let Δ be a set of Σ sentences, and let ϕ be a Σ sentence. We say that Δ **semantically entails** ϕ, written $\Delta \vDash \phi$, just in case ϕ is true in all models of Δ. That is, if v is a model of Δ, then $v(\phi) = 1$.

EXERCISE 1.3.8 Show that if $\Delta, \phi \vDash \psi$, then $\Delta \vDash \phi \to \psi$.

EXERCISE 1.3.9 Show that $\Delta \vDash \phi$ if and only if $\Delta \cup \{\neg\phi\}$ is inconsistent. Here $\Delta \cup \{\neg\phi\}$ is the theory consisting of $\neg\phi$ and all sentences in Δ.

We now state three main theorems of the metatheory of propositional logic.

THEOREM 1.3.10 (Soundness) *If $\Delta \vdash \phi$, then $\Delta \vDash \phi$.*

The soundness theorem can be proven by an argument directly analogous to the substitution theorem in Section 1.4. We leave the details to the reader.

THEOREM 1.3.11 (Completeness) *If $\Delta \vDash \phi$, then $\Delta \vdash \phi$.*

The completeness theorem can be proven in various ways. In this book, we will give a topological proof via the Stone duality theorem (see Chapter 3).

THEOREM 1.3.12 (Compactness) *Let Δ be a set of sentences. If every finite subset Δ_F of Δ is consistent, then Δ is consistent.*

The compactness theorem can be proven in various ways. One way of proving it – although not the most illuminating – is as a corollary of the completeness theorem. Indeed, it's not hard to show that if $\Delta \vdash \phi$, then $\Delta_F \vdash \phi$ for some finite subset Δ_F of Δ. Thus, if Δ is inconsistent, then $\Delta \vdash \bot$, hence $\Delta_F \vdash \bot$ for a finite subset Δ_F of Δ. But then Δ_F is inconsistent.

DEFINITION 1.3.13 A theory T, consisting of axioms Δ in signature Σ, is said to be **complete** just in case Δ is consistent and for every sentence ϕ of Σ, either $\Delta \vDash \phi$ or $\Delta \vDash \neg\phi$.

https://doi.org/10.1017/9781316275603.003 Published online by Cambridge University Press

Be careful to distinguish between the completeness of our proof system (which is independent of any theory) and completeness of some particular theory T. Famously, Kurt Gödel proved that the theory of Peano arithmetic is incomplete – i.e., there is a sentence ϕ of the language of arithmetic such that neither $T \vdash \phi$ nor $T \vdash \neg\phi$. However, there are much simpler examples of incomplete theories. For example, if $\Sigma = \{p, q\}$, then the theory with axiom $\vdash p$ is incomplete in Σ.

DEFINITION 1.3.14 Let T be a theory in Σ. The **deductive closure** of T, written $\mathrm{Cn}(T)$, is the set of Σ sentences that is implied by T. If $T = \mathrm{Cn}(T)$, then we say that T is **deductively closed**.

Example 1.3.15 Let $\Sigma = \{p\}$, and let $T = \{p\}$. Let $\Sigma' = \{p, q\}$, and let $T' = \{p\}$. Here we must think of T and T' as different theories, even though they consist of the same sentences – i.e., $T = T'$. One reason to think of these as different theories: T is complete, but T' is incomplete. Another reason to think of T and T' as distinct is that they have different deductive closures. For example, $q \vee \neg q$ is in the deductive closure of T', but not of T.

The point here turns out to be philosophically more important than one might think. Quine argued (correctly, we think) that choosing a theory is not just choosing axioms, but axioms in a particular language. Thus, one can't tell what theory a person accepts merely by seeing a list of the sentences that she believes to be true. ⌟

EXERCISE 1.3.16 Show that the theory T' from the previous example is not complete.

EXERCISE 1.3.17 Show that $\mathrm{Cn}(\mathrm{Cn}(T)) = \mathrm{Cn}(T)$.

EXERCISE 1.3.18 Consider the signature $\Sigma = \{p\}$. How many complete theories are there in this signature? (We haven't been completely clear on the identity conditions of theories and, hence, on how to count theories. For this exercise, assume that theories are deductively closed, and two theories are equal just in case they contain exactly the same sentences.)

1.4 Translating between Theories

Philosophers constantly make claims about relations between theories – that they are equivalent, or inequivalent or one is reducible to the other, or one is stronger than another. What do all these claims mean? Now that we have a formal notion of a theory, we can consider how we might want to represent relations between theories. In fact, many of the relations that interest philosophers can be cashed out in terms of the notion of a **translation**.

There are many different kinds of translations between theories. Let's begin with the most trivial kind of translation – a change of notation. Imagine that at Princeton, a scientist is studying a theory T. Now, a scientist at Harvard manages to steal a copy of the Princeton scientist's file, in which she has been recording all the consequences of T. However, in order to avoid a charge of plagiarism, the Harvard scientist runs a “find and replace” on the file, replacing each occurence of the propositional constant p with the propositional constant h. Otherwise, the Harvard scientist’s file is identical to the Princeton scientist’s file.

What do you think: is the Harvard scientist’s theory the same or different from the Princeton scientist’s theory?

Most of us would say that the Princeton and Harvard scientists have the same theory. But it depends on what we mean by “same.” These two theories aren’t the same in the strictest sense, since one of the theories contains the letter “p,” and the other doesn’t. Nonetheless, in this case, we’re likely to say that the theories are the same in the sense that they differ only in ways that are incidental to how they will be used. To borrow a phrase from Quine, we say that these two theories are **notational variants** of each other, and we assume that notational variants are equivalent.

Let’s now try to make precise this notion of “notational variants” or, more generally, of **equivalent theories**. To do so, we will begin with the more general notion of a translation from one theory into another.

DEFINITION 1.4.1 Let Σ and Σ' be propositional signatures. A **reconstrual** from Σ to Σ' is a function from the set Σ to the set $\mathsf{Sent}(\Sigma')$.

A reconstrual f extends naturally to a function $\overline{f} : \mathsf{Sent}(\Sigma) \to \mathsf{Sent}(\Sigma')$, as follows:

1. For p in Σ, $\overline{f}(p) = f(p)$.
2. For any sentence ϕ, $\overline{f}(\neg\phi) = \neg\overline{f}(\phi)$.
3. For any sentences ϕ and ψ, $\overline{f}(\phi \circ \psi) = \overline{f}(\phi) \circ \overline{f}(\psi)$, where $\circ$ stands for an arbitrary binary connective.

When no confusion can result, we use f for $\overline{f}$.

THEOREM 1.4.2 (Substitution) *For any reconstrual $f : \Sigma \to \Sigma'$, if $\phi \vdash \psi$ then $f(\phi) \vdash f(\psi)$.*

Proof Since the family of sequents is constructed inductively, we will prove this result by induction.

(rule of assumptions) We have $\phi \vdash \phi$ by the rule of assumptions, and we also have $f(\phi) \vdash f(\phi)$.

($\wedge$ intro) Suppose that $\phi_1, \phi_2 \vdash \psi_1 \wedge \psi_2$ is derived from $\phi_1 \vdash \psi_1$ and $\phi_2 \vdash \psi_2$ by $\wedge$ intro, and assume that the result holds for the latter two sequents. That is, $f(\phi_1) \vdash f(\psi_1)$ and $f(\phi_2) \vdash f(\psi_2)$. But then $f(\phi_1), f(\phi_2) \vdash f(\psi_1) \wedge f(\psi_2)$ by $\wedge$ introduction. And since $f(\psi_1) \wedge f(\psi_2) = f(\psi_1 \wedge \psi_2)$, it follows that $f(\phi_1), f(\phi_2) \vdash f(\psi_1 \wedge \psi_2)$.

($\to$ intro) Suppose that $\theta \vdash \phi \to \psi$ is derived by conditional proof from $\theta, \phi \vdash \psi$. Now assume that the result holds for the latter sequent, i.e., $f(\theta), f(\phi) \vdash f(\psi)$. Then conditional proof yields $f(\theta) \vdash f(\phi) \to f(\psi)$. And since $f(\phi) \to f(\psi) = f(\phi \to \psi)$, it follows that $f(\theta) \vdash f(\phi \to \psi)$.

https://doi.org/10.1017/9781316275603.003 Published online by Cambridge University Press

(reductio) Suppose that $\phi \vdash \neg\psi$ is derived by RAA from $\phi, \psi \vdash \bot$, and assume that the result holds for the latter sequent, i.e., $f(\phi), f(\psi) \vdash f(\bot)$. By the properties of f, $f(\bot) \vdash \bot$. Thus, $f(\phi), f(\psi) \vdash \bot$, and by RAA, $f(\phi) \vdash \neg f(\psi)$. But $\neg f(\psi) = f(\neg\psi)$, and, therefore, $f(\phi) \vdash f(\neg\psi)$, which is what we wanted to prove.

($\vee$ elim) We leave this step, and the others, as an exercise for the reader. □

DEFINITION 1.4.3 Let T be a theory in Σ, let T' be a theory in Σ', and let $f : \Sigma \to \Sigma'$ be a reconstrual. We say that f is a **translation** or **interpretation** of T into T', written $f : T \to T'$, just in case:

$$T \vdash \phi \quad \Longrightarrow \quad T' \vdash f(\phi).$$

Note that we have used the word "interpretation" here for a mapping from one theory to another, whereas we previously used that word for a mapping from a theory to a different sort of thing, viz. a set of truth values. However, there is no genuine difference between the two notions. We will soon see that an interpretation in the latter sense is just a special case of an interpretation in the former sense. We believe that it is a mistake to think that there is some other (mathematically precise) notion of interpretation where the targets are concrete (theory-independent) things.

DISCUSSION 1.4.4 Have we been too liberal by allowing translations to map elementary sentences, such as p, to complex sentences, such as $q \wedge r$? Could a "good" translation render a sentence that has no internal complexity as a sentence that does have internal complexity? Think about it.

We will momentarily propose a definition for an equivalence of theories. However, as motivation for our definition, consider the sorts of things that can happen in translating between natural languages. If I look up the word "car" in my English–German dictionary, then I find the word "Auto." But if I look up the word "Auto" in my German–English dictionary, then I find the word "automobile." This is as it should be – the English words "car" and "automobile" are synonymous and are equally good translations of "Auto." A good round-trip translation need not end where it started, but it needs to end at something that has the *same meaning* as where it started.

But how are we to represent this notion of "having the same meaning"? The convicted Quinean might want to cover his eyes now, as we propose that a theory defines its own internal notion of sameness of meaning. (Recall what we said in the preface: that first-order metatheory is chalk full of intensional concepts.) In particular, ϕ and ψ have the same meaning relative to T just in case $T \vdash \phi \leftrightarrow \psi$. With this notion in mind, we can also say that two translations $f : T \to T'$ and $g : T \to T'$ are synonymous just in case they agree up to synonymy in the target theory T'.

DEFINITION 1.4.5 (equality of translations) Let T and T' be theories, and let both f and g be translations from T to T'. We write $f \simeq g$ just in case $T' \vdash f(p) \leftrightarrow g(p)$ for each atomic sentence p in Σ.

With this looser notion of equality of translations, we are ready to propose a notion of an equivalence between theories.

https://doi.org/10.1017/9781316275603.003 Published online by Cambridge University Press

DEFINITION 1.4.6 For each theory T, the identity translation $1_T : T \to T$ is given by the identity reconstrual on Σ. If $f : T \to T'$ and $g : T' \to T$ are translations, we let gf denote the translation from T to T given by $(gf)(p) = g(f(p))$, for each atomic sentence p of Σ. Theories T and T' are said to be **homotopy equivalent**, or simply **equivalent**, just in case there are translations $f : T \to T'$ and $g : T' \to T$ such that $gf \simeq 1_T$ and $fg \simeq 1_{T'}$.

EXERCISE 1.4.7 Prove that if v is a model of T', and $f : T \to T'$ is a translation, then $v \circ f$ is a model of T. Here $v \circ f$ is the interpretation of Σ obtained by applying f first, and then applying v.

EXERCISE 1.4.8 Prove that if $f : T \to T'$ is a translation, and T' is consistent, then T is consistent.

https://doi.org/10.1017/9781316275603.003 Published online by Cambridge University Press