Principles of mathematics and logic
A course for Liberal Arts students

R.A.G. Seely
Mathematics Department
John Abbott College
Ste Anne de Bellevue, QC

Version 2020.9.15

Contents

Note to students

This book is intended to accompany the *Principles of Mathematics and Logic* course, given in the Liberal Arts program at John Abbott College. This course covers virtually all the material in the text; you should expect to read it cover to cover. Of course, there are also lectures which make up an important part of the course; you will find that often I emphasise things somewhat differently in class and in the book—the intention is that each should complement the other, rather than replace it. You should not skip class, expecting to make it up with the text (instead, attend class regularly), and similarly, you should not rely solely on your class notes (read the book for the extra examples and explanations). The most important part of the book is the exercises: it is a (true!) cliché that mathematics is a poor spectator sport, and to *learn* mathematics properly, you must *do* mathematics. Take this seriously: you will find it very hard to succeed unless you actually practice the ideas learned in class.

Generally, when you read this text, indeed any mathematics, it is important to engage the text actively, not passively. You should have pencil and paper beside you, and try to follow each statement, doing the suggested calculations or reasoning yourself. It is not a novel or short story, whose meaning will just flow over you, but a dialogue, only one side of which is on the page. You must provide the other side yourself!

In particular, you will find lots of examples with explanations; try to do the examples yourself (especially after the first one, or after seeing some in class). A good idea is to try to do an example without looking at the explanation, only turning to the text for hints as you go. When you've done the example, read through my explanation to see if you understand it all, and then go onto the next example. And of course, *do the exercises!*

There is a course webpage (`www.math.mcgill.ca/rags/jac.html`); I have put additional material there, including further readings (some intended to give you further explanation and examples of topics covered in class, some intended to go further in some topics than covered by the course, and some intended to interest you, without any intent at "examinable material"), further exercises (particularly practice tests to help you prepare for the class tests), and any other relevant information (for instance, your marks will be found there after tests). You should bookmark the webpage, and visit it often to see if I've put new material there for you.

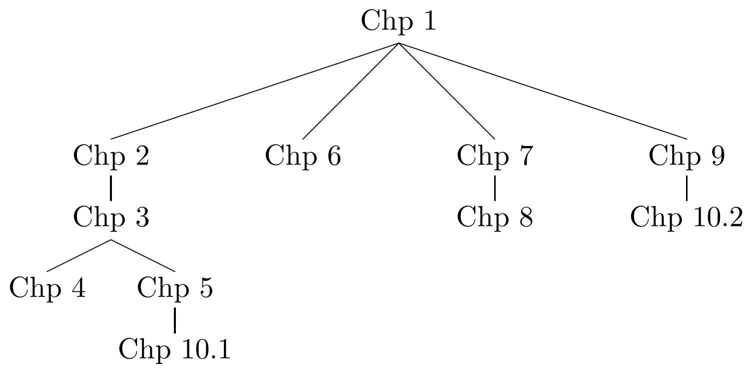
Acknowledgements

This book is based on many sources, in addition to my own work: I am most particularly indebted to the earlier text written by the late Gerald LaValley, used in this course for many years. A substantial portion of this text is inspired by Gerry's book, especially Chapters 1 and 6–8 which are heavily influenced by his approach. Gerry also carefully read early drafts of my text, and I am very grateful for his corrections, comments and suggestions. I have also benefited from notes used at the University of Ottawa, written by Phil Scott and Peter Selinger, and from *Formal Systems and Logic in Computing Science* by W.W. Armstrong, F.J. Pelletier, R.A. Reckhow, & P. Rudnicki. In addition, the section on Knights and Knaves (Chapter 1) relies heavily on the related book by Raymond Smullyan.

Many decades ago I learned from my brother John that the right way to combine the Natural Sciences with the Liberal Arts is with a conjunction, not with a disjunction. I hope to help the reader to get this message too: engage deeply with at least one aspect of the human story, but remain in touch with the rest of the tale. Otherwise you will not be as complete a person as you can be.

Major dependencies

The text is intended to be read (and covered in class) in order, but in fact some chapters only depend lightly on some prior chapters. So, with minor dependencies that can be "ignored", here is the essential dependency graph: you really need the chapters above a certain chapter to understand the latter.



Note: From Chapter 7 onwards, the language of set theory is used, but little of the "serious" content of Chapter 6 is needed. So, these later chapters do depend on Chapter 6, but only lightly. Section 10.2 is a bit unusual: it really does not use previous material in a serious way, but it will perhaps make more sense to a reader familiar with what's gone on earlier in the text. I've indicated this with a dependency on Chapter 9, but even that isn't really true(!).

Section 10.1 depends to some extent on understanding the basic structure of the natural numbers, so in a sense also depends on Chapter 8, but could be read without that. For example, mathematical induction (from Chapter 8) is mentioned, but not used in any essential manner.

For John, in memory

Chapter 1

Introduction to Logic

1.1 Mathematics and Science in A Liberal Arts Education

Historically the liberal arts included arithmetic, geometry, astronomy, fine arts and history, grammar, rhetoric and logic. As broad general education, Liberal Arts programs are alternatives to training in a trade or craft.

Some students think of Liberal Arts as the history of (Western) culture and the themes, styles and movements in literature and the fine arts—an encyclopedia of cultural facts; lists of historical particulars.

General (liberal arts) education has always been more than just particular truths in a narrow range of fields. It aims to reveal the relevance of these truths, the connections and relations among the particulars, and the subsumption of particulars under abstract general principles. Students should understand not just the truths but the search for truth; not just knowledge but the methods by which we acquire and confirm knowledge. Facts are important, but the interpretation of what the facts mean is crucial.

Logic is central to this understanding.

Western culture is the result of developments in mathematics and the physical and social sciences as much as it is a product of "merely historical" accidents or of changes in artistic or literary directions. Human creativity and awareness are as evident in logic, mathematics, and the sciences as they are in philosophy and the fine arts. For those who develop the understanding, sensitivity, and taste, the great logical and mathematical proofs and the deep and subtle theories of the sciences are as beautiful and as admirable as any product of the human spirit.

1.1.1 Logic and rationalization

"Logic" has been defined as the science of right reasoning.

Freud and Marxists and the existentialists encourage a common confusion about the relationship between logic, rationality, and rationalization. Their idea is that people use logic to "explain away" behaviours and attitudes whose real explanations are non-logical. Freudians claim that one's beliefs are not grounded on logical reasons but have their source in the sub-conscious; Marxists blame ideology; existentialists emphasize "bad faith".

The science of logic begins with a value-judgment: "what is right reasoning?". The identification of logic with rationalization ("explaining away") is based on a relativistic view of values. The claim is that there is no one standard of right reasoning, but that "right" (like any value) is a matter of taste. Standard canons of logic are decided by whatever social group (class, gender, etc.) has the power to impose its standards of "right reasoning" on the rest of society. So Marxists claim that logic is a bourgeois requirement. Some feminists claim that it is something that males impose on the world.

This text rejects such relativism.[1]

1.1.2 Many logics, not just one logic

May one admit different standards of "right reasoning" if one is considering different contexts? It seems perfectly reasonable that one might. However, admitting that there are various notions of "right reasoning" does not mean that one admits that the notion of "right reasoning" is merely a matter of taste. One thing we insist on is what philosophers of science call "reproducibility": if two reasonable observers observe the same phenomenon, they will make the same observation. We shall insist, then, that "right reasoning", and the logic that encodes it, must satisfy this requirement of "reproducibility"; logic is no mere matter of opinion.

This may well make developing a logic suitable for political purposes an impossible task! In fact, what logic is suitable for a specific purpose may well be a matter of opinion (and often is!). But that gets us into philosophical disputes that have lasted for centuries, and so takes us way beyond what we can cover in one semester.

So, it has to be said right at the start of the course that there is no single logic which encompasses the idea of capturing "right reasoning"; instead there are several candidates which have been studied in considerable detail in the past century. Each of these logics attempts to capture specific aspects of right reasoning, usually suitable for specific applications or circumstances. We shall emphasize (mainly for reasons of simplicity) the traditional "classical" logic, whose origins go back to the Greek philosophers (such as Aristotle), and which was what was principally meant by the term "logic" for centuries.

A very important variant of classical logic, which became a serious matter of philosophical and mathematical scrutiny only around the early 20^{th} century, is "intuitionist logic", a logic intended to capture a more "constructive" aspect of logic; we shall discuss this at several junctures as we study classical logic. With the development of computing science, an increasing need grew for a logic which was more "resource sensitive"; this need is met by a collection of what are known as "sub-structural logics". These were not new to philosophers, as they had already been considering various "relevance logics", which were an attempt to address some seemingly paradoxical behaviour of the classical notion of *implication* (remarked upon near the end of Section 1.3.2); in Chapter 10 we shall briefly consider how relevance logic can avoid some of this. Related to such logics are various candidates for a logic of quantum phenomena, a logic suitable for underlying (for example) quantum computing, and more generally for understanding quantum physics. This is a field of current active research.

Other distinctions are possible: for instance there is a large family of "modal logics",[2] whose intention is to study the logical properties of notions like "possibility", "necessity", "belief", *etc.* Again, we shall discuss these, but only briefly. Another distinction is made between deductive and inductive logic (inductive logic is very commonly used in the natural sciences, less so in mathematics): this is essentially a distinction between a logic aiming at determinate, definite conclusions and a logic aiming at probabilistic conclusions. Inductive logic is often identified with statistics, but there is active current research into a suitable formal logic for such matters.

[1] Perhaps this is as good a time as any to mention that mid-way through the course, you will be asked to read the essay *On Bullshit* by Harry Frankfurt.

[2] In the wide sense, "modal logic" includes many logics such as temporal, deontic, and doxastic logic, for instance. You may Google these, if you wish, as we shall not go there!

Furthermore, even traditional logics (such as the classical logic we shall study) may be enhanced by adding other features, to allow other aspects of "right reasoning" to be captured. For instance, we may add to classical logic the ability to make sentences of the form "infinitely many objects have such-and-such a property" (in addition to the sentences of the form "all objects have such-and-such a property" and "some object has such-and-such a property", which are already allowed, and which we shall study). The list of variants seems almost endless (in a sense, it is!).

Many logics for many purposes—each has its own characteristics, its own properties. The study of most of these logics follows a similar plan, which is simplest in the case of classical logic, so we shall use that as an illustration of how a formal logic, a logic to capture the elusive notion of "right reasoning", may be developed in a scientific, perhaps we should say mathematical, fashion.

1.1.3 Pure mathematics and logic

In this course we mostly study pure (or formal or theoretical) mathematics and logic, more than applied mathematics and logic. We study maths and logic from a theoretical point of view. Practical applications (such as using logic to persuade somebody or using mathematics for utilitarian calculations) are secondary. The aim is to develop some understanding of what these two sciences are about, of their methods, and of their beauty and interest for their own sakes.

So we study the principles of these two fields of study, not how one applies them. In logic, the course does not aim to teach rhetoric. In mathematics, this means that our ability to perform calculations will not be emphasized. Unlike most college mathematics courses, this course does not emphasize applying the techniques of trigonometry and differential and integral calculus. We look at the basic assumptions behind the two fields, the way mathematicians and logicians arrive at their basic assumptions, and the way they arrive at conclusions based on those assumptions. In particular, we do *not* regard mathematics as "the science of quantity", or any such definition (if this challenges your preconceptions, so much the better!). Rather, we regard the essence of mathematics (including logic) as the study of *pattern*. The word "pattern" means many things of course, but one property that is intended by my usage is "reproducibility": whether or not a pattern is (say) beautiful is (probably) not a reproducible property (we may disagree on whether something is beautiful), but the pattern itself is—by the very nature of what a pattern is. Mathematicians study pattern in many contexts: among numbers, geometric shapes, and logic, to be sure, but also in other domains, including (but not limited to) music, natural language, computer programming languages and computer programs themselves, as well as more "useful" domains, like the movement of planetary bodies, and the performance of the stock exchange. In this course we'll see a few instances of this, at an elementary level: we'll consider patterns in logic, in numbers and sets, and (time permitting) in natural language.

In addition, since the early twentieth century, mathematics (especially pure mathematics) has tended to have a characteristic method or procedure, often referred to as "the axiomatic method" (we shall study some examples at the end of the course). In studying a subject or discipline, one begins with a set of undefined elements, properties and relations among these elements, and fundamental "truths" called postulates or axioms, which establish the basic facts of the subject. From these all other facts (theorems) should be derived by formal logic, without appeal to any external knowledge. Some commentators stress that the undefined elements should not be regarded as concrete entities, but rather as some sort of "variables", which may be interpreted in any way consistent with the axioms—in this way, the subject becomes merely the study of what consequences may be derived from the initial axioms. It might appear that this approach tends to identify mathematics with "applied logic", and indeed, there was a serious attempt to reduce mathematics to logic early in the twentieth century with the mammoth 3-volume set *Principia Mathematica*

(Cambridge University Press, 1910) by A.N. Whitehead and B. Russell.

But the reduction of mathematics to logic suffered serious blows, right from its inception as an idea. Firstly, when Whitehead and Russell tried to implement their idea, they found they needed a non-logical axiom of infinity in order to even capture simple arithmetic, in spite of many efforts to avoid such an extra assumption. (They could describe simple "natural numbers" like 1 and 2 in purely logical terms, and they could even, after several hundred pages, prove that $1 + 1 = 2$, but what they could not do was talk about *all* natural numbers without the axiom of infinity.)

Then, in the early 1930s, Kurt Gödel proved that in their system (or in any similar system for mathematics) there were statements (which were "obviously true" in some sense) which could never be proven nor disproven (unless their system was in fact inconsistent). Gödel explicitly saw this result of his as showing that mathematics could never be reduced to a merely formal or logical system, but that some other considerations, mathematical considerations, were an essential part of the story. Moreover, he believed that mathematics studied real phenomena, not merely intellectual creations: entities such as numbers, geometric shapes, *etc.*, may not exist as tangible objects like rocks, cats and dogs, or even MP3 files, but they have a reality nonetheless. This view is often called "Platonic" (for reasons I need not explain to this audience!). We may discuss these matters when we consider Gödel's theorem.

Furthermore, while the axiomatic method certainly describes the modern practice of a considerable body of mathematics, and it does indicate the close relationship between mathematics and logic (a relationship we shall see throughout the course), nonetheless one must remember to distinguish between 'method' and 'essence'. The method does describe part of what mathematicians do (and how they do it), but still it does not entirely address the essence of what mathematics *is*, a much more complicated and obscure matter. In particular, it ignores the question of what makes some collections of axioms more valuable as an object of study than others, for instance. That usually involves the consideration of what the mathematics is used for, whether for other parts of mathematics or for "real-life problems"; it may also involve matters of "taste" (and the less reputable, but equally compelling, notion of "fashion"), and often simply what captures the imagination and passion of the mathematicians working in a discipline. One key motivation is the love of beauty (which brings us back to "pattern").

1.1.4 Patterns in sciences and arts

For a Liberal Arts student, the main relation between the sciences on one hand and the fine arts and literature on the other is that both study patterns (there's that word again!). Empirical sciences study the patterns in nature. Logic studies the patterns of human reasoning. Mathematics studies the patterns to be found in patterns.

It shares many characteristics with music in this aspect of its nature, an observation made often by many writers and mathematicians. In a paper on Newton, the mathematician James Joseph Sylvester (himself a talented amateur musician) wrote

> May not Music be described as the Mathematics of sense, Mathematics as the Music of reason? Thus the musician feels Mathematics, the mathematician thinks Music—Music the dream, Mathematics the working life—each to receive its consummation from the other when the human intelligence, elevated to its perfect type, shall shine forth glorified in some future Mozart-Dirichlet or Beethoven-Gauss.[3]

[3] Quoted in Edward Rothstein, *Emblems of Mind: The Inner Life of Music and Mathematics*, Times Books 1995. Sylvester (1814-1897) was an English algebraist who spent his professional life in both the US and England. Mozart and Beethoven need no introduction; Dirichlet (1805-1859) was a German mathematician active in the field of analysis.

Æsthetics and appreciation of the arts involves feeling and responding to the patterns in nature and in works of art. Artistic creation is a matter of creating or reproducing or elaborating patterns. Empirical science is the discovery, description and analysis of patterns in nature. The formal sciences (logic and mathematics) study, describe, and create patterns of patterns. The beauty of structure and pattern is as central to the study of logic and mathematics as it is to literature, music or painting. Some recognition of this centrality is the main thing I hope students will develop through this course.

1.2 Introduction to Logic

1.2.1 Some history

A *very* brief, even superficial, history of logic may help put the content of this course into some context.[4]

The western scholarly study of logic goes back to Aristotle, who listed the variants of the syllogism. The law of non-contradiction and the law of the excluded middle are also credited to Aristotle. Most European students of logic followed the Aristotelian tradition until the mid nineteenth century.

However, another classical Greek school of logic, which may be loosely identified with the Stoics, went beyond the syllogistic tradition, and essentially had an understanding of propositional logic, in almost the modern sense. They formulated five basic "inference schemata":

- If the first, then the second; but the first; therefore the second.
- If the first, then the second; but not the second; therefore, not the first.
- Not both the first and the second; but the first; therefore, not the second.
- Either the first or the second [and not both]; but the first; therefore, not the second.
- Either the first or the second; but not the second; therefore the first.

We'll see how these fit into our presentation of logic at the end of Chapter 2.[5]

An early precursor to the mathematical tradition in logic was Leibniz, who envisioned a calculus of logic, a set of rules which would completely automate the reasoning process, so that disagreements might be settled by simple calculation. He even imagined one might build a machine to do these calculations mechanically. Needless to say, this idea (dare I call it a dream?) of Leibniz's was never realized in his lifetime.[6]

But in the nineteenth century, an algebraic approach to propositional logic was successfully designed by George Boole (we shall study Boolean algebras in the last section of the course). This

---

Curiously, given the nature of this quotation, he was Felix Mendelssohn's brother-in-law. Gauss (1777-1855) was one of the greatest mathematicians ever, who contributed significantly to just about every aspect of mathematics, as well as to physics (he was particularly famous during his lifetime for his contributions to the understanding of electricity and magnetism, and for his part in developing the telegraph).

[4]A slightly longer account, from the *Oxford Companion to Philosophy*, may be found on my webpage.

[5]But for now notice that in the 4th of these schemata the parenthetical phrase "and not both" renders the scheme rather redundant. The Stoics would not have had that phrase, and in effect, this really just says that (unlike modern logicians) they took "or" to be the "exclusive or", which we'll define very soon.

[6]Leibniz did get quite close to the idea behind the system Boole developed in one remark he made, using numbers to represent "thoughts" or properties: "if the term for an 'animate being' should be imagined as expressed by the number 2, and the term for 'rational' by the number 3, then the term for 'man' will be expressed by the number 2×3, that is 6". [Quoted in Kitty Ferguson, *The Music of Pythagoras*, Walker Books 2008.]

could be marked as the first step in modern logic. What was missing, however, was a mathematical account of predicates, and what we call quantifiers (Chapter 5 of this text). That was managed by Gottlob Frege later in the century, in a remarkable text called *Begriffsschrift*, whose unconventional notation makes this text an effort to penetrate for most readers. His ideas were quickly picked up by the mathematical-logical community, however, and with a notation very like the one we use in this text, became the basis for twentieth century logic. Very soon after, Peano gave an axiomatization of the theory of natural numbers, and although there are technical reasons why a complete mechanization of the rules of predicate logic cannot exist, one could claim the essence of Leibniz's dream was realized. (One of those technical reasons is Gödel's incompleteness theorems, which we shall study at the end of the course.)

In the late nineteenth, early twentieth century, logic went hand in hand with the attempt to put the foundations of mathematics on a sure foundation. A number of paradoxes were becoming apparent in mathematics, especially with the study of the infinite, and a need for a firm philosophical basis for mathematics and logic was thought to be necessary. Set theory was a central tool in this attempt, but a series of paradoxes in logic and set theory underlined just how conceptually tricky ("subtle" might be a more positive term) things were.

The paradoxes

Without getting too technical, let's consider some of the paradoxes that caused such a concern. One is very old, in fact: it is often called *The Liar* and finds its origins in classical Greece. Consider the statement: "This statement is false". If it's false, it's true, but if it's true, it's false. There are many variants of this; here is a simple one. Imagine a card with the following two statements, one on one side, the other on the other: "The statement on the other side of this card is false", "The statement on the other side of this card is true".

The work of Frege was interrupted by Bertrand Russell, who found an error in his system, which allowed sets to be formed if defined by some property. The paradox Russell found was this (if your knowledge of sets is insufficient, come back to this story after we've done chapter 6—in any case I'll give a simpler version in a moment): notice that some sets seem to contain themselves as an element, such as the set of abstract entities (it is itself an abstract entity), whereas other sets (most, in fact) do not, such as the set of words on this page, which is itself *not* a word on the page. We'll call those sets which do *not* contain themselves as an element "standard sets", and those which *do* contain themselves as an element "non-standard sets". Consider now the collection of all standard sets: is this set standard or not? If it's standard, then it's non-standard, but if it's non-standard, then it's standard.

If this is too technical, here's a non-technical variant. Consider a village, with just one (male) barber, who shaves every man in the village who does not shave himself, and no one else. Who shaves the barber? If he shaves himself, then he cannot shave himself, but if he doesn't shave himself, then he does shave himself.

Here's a numerical paradox: notice that some numbers may be described with only a few words ("one", "the hundredth prime"), and others take more words ("one million seven hundred and forty five thousand three hundred and twenty nine"). Generally (though there are exceptions), numbers that take lots of words tend to be large, and ones that can be described in fewer tend to be smaller. Here is an interesting number: the smallest number that cannot be described in less than thirteen words. What's the paradox? I just described it in twelve words.

There are oodles of other paradoxes—what they all have in common is that trying to understand them caused mathematicians and philosophers to think hard about mathematics and logic, and that resulted in a clearer understanding of what is going on in those domains. Many of these paradoxes still inspire thought and commentary (though for the practicing mathematician, they are more like pleasant entertainments these days).

One serious result of the philosophical ferment in the early decades of the twentieth century was the development of an alternate view of mathematics and logic, which goes by the general name of intuitionism. We shall consider some aspects of intuitionism later in the course, but for now, let's just say that it demanded a more constructive approach to mathematics and logic. For instance, if you asserted the existence of something, in intuitionist practice, you had to actually show the thing in question, or give a clear algorithm for its construction. Here is an example (though you may have to return to this after we study Chapter 7 to understand some of the terms): suppose you wonder if there are numbers x, y with the property that x, y are not rational (are not expressible as fractions of natural numbers), and yet x^y is rational. Here is one possible answer, one that would have been accepted by most mathematicians at the end of the nineteenth century (and would still be acceptable today by most—non-intuitionist—mathematicians): consider $\sqrt{2}^{\sqrt{2}}$. Either this is rational or it is not. If it is rational, then since $\sqrt{2}$ is not rational, you have your numbers ($x = y = \sqrt{2}$). If it isn't rational, then again you have your numbers: just take $x = \sqrt{2}^{\sqrt{2}}$, which you have supposed isn't rational, and $y = \sqrt{2}$, which isn't rational, but now $x^y = (\sqrt{2}^{\sqrt{2}})^{\sqrt{2}} = \sqrt{2}^{(\sqrt{2}\sqrt{2})} = \sqrt{2}^{\,2} = 2$ which is rational. What's wrong with this, according to the intuitionists? Simply that at the end of the argument, you still don't really know what values x, y are that have the required property. Just what is x? You don't know from this argument: it might be $\sqrt{2}$ or it might be $\sqrt{2}^{\sqrt{2}}$—the argument didn't specify which one it really was. Your proof wasn't constructive, in that it didn't put the necessary values at your fingertips.[7] During the past century, intuitionist logic has had a lot of study, and has become very important for practical reasons, for example in theoretical computer programming, where constructivity is a key ingredient, as well as for philosophical purposes.

What about logic today? The past century has been (and continues to be) a golden age in mathematical logic, with ever more impressive gains in understanding and in practical and theoretical applications of that understanding to many disciplines. There are several features of contemporary mathematical logic that distinguish it from the past practices. Perhaps the most striking is that one no longer thinks of "logic" as a single entity, but rather there are many different "logic**s**", for many different purposes. Logic(s) is(are) studied with mathematical tools, and indeed, logics are mathematical entities in their own right. We shall see a simple example of this at the end of the course, when we consider a logic suitable for the analysis of sentence structure in linguistics.

1.2.2 Some vocabulary

Logic is the science of discursive reasoning.[8]

As a science, logic aims to discover general laws that apply to all discursive reasoning. Narrowly specific kinds of reasoning that are only relevant to some particular subject matter are the concern of the special sciences.

Discursive reasoning consists of arguments made up of statements.

A **statement** is made by a declarative sentence. A statement is either true or false (although its truth or falsity may not be known). No statement is both true and false.

An **argument** is a collection of one or more statements called **premises** and one statement called the **conclusion**. Premises are offered as grounds for the conclusion.

[7] Actually, one can give a constructive proof. For example, it's a fact that $\sqrt{5}$ and $\log_5 9$ are irrational (these facts follow from the results of Chp 8), but (using your high-school algebra!) $\sqrt{5}^{\log_5 9} = (5^{\log_5 9})^{1/2} = \sqrt{9} = 3$.

[8] This is a provisional definition—we shall improve on it as we go.

Premise-statements are grounds for a conclusion when the truth of those premise-statements gives some assurance that the conclusion-statement is true. When the premises (if true) really do give some assurance that the conclusion is true, we say that the premises support the conclusion, or that they **entail** the conclusion, or that the conclusion follows from the premises. In such cases, we say the argument is **valid**. If the conclusion does not follow from the premises, we call the argument **invalid**.

Whether a collection of statements is an argument depends on the intention of the person who makes the statements. It is an argument if she intends some of the statements as grounds for a statement that she offers as a conclusion. If the premises do not support the conclusion, it is still an argument, but it is a bad (invalid) argument. If the person never intended the premises to support or entail the conclusion, it is not an argument.

In logic, we do not use "valid" to describe a statement. *Statements* are true or false; *arguments* are valid or invalid.

Using this vocabulary, we refine the definition of "logic" given above. **Logic** is the science that studies the general principles or laws of valid arguments.

Deductive logic is the science of **deductive arguments**. In a good deductive argument (a valid deduction), the conclusion *cannot be false* if the premise(s) are true. The rules for deductive reasoning guarantee that one cannot get a false conclusion from true premises.

Inductive arguments offer less assurance than deductive arguments. In a good inductive argument (a valid induction), true premises assure us only that the conclusion is *probably true*. A valid inductive argument makes it *rational to believe* that its conclusion is true, while allowing that it might turn out to be false.[9]

Most of this course (and most of mathematics and logic) concerns deductive arguments. Inductive argument is touched on in the section on the mathematics of probability and statistics.

1.2.3 Beginning steps in deductive logic

What kinds of arguments *guarantee* that their conclusions cannot be false when their premise(s) are true?

An obvious example of such an argument is an argument based on definitions. For example, if we define "bachelor" as an adult unmarried human male, we could argue:

John is a bachelor.	(Premise)
(Therefore) John is unmarried.	(Conclusion)

The conclusion cannot be false if the premise is true, because "bachelor" means (among other things) "unmarried".

Another obvious example is the classic:

All men are mortal.	(Premise)
Socrates is a man.	(Premise)
(Therefore) Socrates is mortal.	(Conclusion)

This argument is an example of *predicate logic*.

Before looking at predicate logic, we study *propositional logic*. Propositional logic studies arguments whose conclusions depend on the way compound statements are composed of simple statements and special "connectives". The compound statement "Ivanhoe is safe and Rebecca is

[9]These definitions of "deductive" and "inductive" are (demonstrably) better than most definitions found in dictionaries—even *good* dictionaries. Dictionaries are not the ultimate arbiters of meaning.

relieved" consists of two *simple* statements ("Ivanhoe is safe", "Rebecca is relieved") linked with the *conjunction connective* "and". "Nero is not pleased" is the negation of the simple statement "Nero is pleased" that results from adding the *negation ("not") connective.*

Here is an example of a propositional logic argument. From the (compound) statement "Ivanhoe is safe and Rebecca is relieved" we can infer "Ivanhoe is safe". That is, in the argument:

Ivanhoe is safe and Rebecca is relieved.	(Premise)
(So) Ivanhoe is safe.	(Conclusion)

when the conjunction "Ivanhoe is safe and Rebecca is relieved" is true, "Ivanhoe is safe" *cannot* be false.

Notice that the validity of the Ivanhoe argument does not depend on the fact that it's about Ivanhoe and Rebecca. Look at the argument:

Mickey Mouse is safe and Minnie is relieved	(Premise)
Therefore Mickey Mouse is safe.	(Conclusion)

This is just as good (valid) as the Ivanhoe argument.

Frodo is an airhead and Arwen is neat.	(Premise)
Frodo is an airhead.	(Conclusion)

Here again the conclusion cannot be false when the premise is true. It's just as good as the Ivanhoe and Mickey arguments.

Dusty is silly but he's beautiful.	(Premise)
Dusty is silly.	(Conclusion)

This is an argument that behaves exactly like the others even though it uses "but" instead of "and" as the connective in the premise. A statement that results from linking two sentences with "but" works the same way as one that uses "and". Both kinds of statements are *conjunctions*. Another word that has the same logical use (*i.e.*, meaning) as "and" and "but" is "while" (as in "I'll wait in the car while you go in". This permits the valid deductive inference "I'll wait in the car"). Others are "whereas" and "as" and "at the same time as" and so on. These are all instances of the same logical connective, *conjunction*: they may have slightly different meanings in ordinary English, but they all have the same property that the truth of the compound sentence requires the truth of the first (indeed, both) constituent parts. This property characterizes conjunction.

As in any science, these observations enable us to *propose a conjecture*. We propose as a *law of deductive logic* (a principle of valid deductive reasoning) that, **given any conjunction as a premise, we may validly infer the first conjunct**. This means that whenever any conjunction is true, its first conjunct cannot be false. We can restate the conjecture as a proposal for a **valid argument form**. We propose that any argument of the form:

Statement 1 and/but/while/whereas ... Statement 2	(Premise)
Statement 1	(Conclusion)

is a valid argument.

You should convince yourself that a similar principle allows the conclusion of Statement 2 from the same premise: check each example to see how "obvious" this is.

Statement 1 and/but/while/whereas ... Statement 2	(Premise)
Statement 2	(Conclusion)

Another example of a valid deductive inference involves a **conditional**, as in the inference:

If you passed logic then I am delighted.	(Premise)
You passed logic.	(Premise)
(Therefore) I am delighted.	(Conclusion)

If the two premises are true, the conclusion cannot be false.

Other English words might serve the same logical purpose as the "if ... then ..." construction; other languages have other words to do the same job. The validity of the argument has nothing to do with your success in the course or with my happiness. We conjecture that this is another valid argument form, expressed as:

If Statement 1 then Statement 2	(Premise)
Statement 1	(Premise)
Statement 2	(Conclusion)

This process of generalization leads us to propose laws of valid arguments that do not depend on whether the premises are actually true. **The validity of an argument only depends on whether the conclusion follows from the premises.**

A deductive argument aims to provide true premises and it aims to provide assurance that the conclusion cannot be false when the premises are true. But a deductive argument might have true premises and a true conclusion even if the premises do *not* guarantee the truth of the conclusion. In such a case, the argument is (deductively) **invalid**. An argument can be valid even if the conclusion and premises are not true.

The proposed rules above did not specify anything about what Statement 1 or Statement 2 were about. The *content* of the statements was not relevant to the validity of arguments. The rules also were quite general about what particular words were used for the connective. In the first conjecture, *any* word in any language was acceptable, as long as it *did the same logical work as* the conjunction-connective "and". I used "and/but/while/whereas..." to indicate that it doesn't matter which of these words was used for the conjunction connective. Similarly, in the second conjecture, other verbal expressions might be used for "if ... then ...", as long as they do the same logical work.

The conjectured rules describe only the **forms** of valid arguments (which is why I said they were "proposals for valid argument **forms**"). We refine our definition of deductive validity[10], to read:

> **A valid deductive argument is an argument whose *form* is such that it is *impossible* to construct an argument *of that form* that has true premises and a false conclusion.**

From this definition, it follows that deductive invalidity can be defined as:

> A deductive argument is **invalid** when **it has a form such that one *could* construct another argument *of the same form* whose premises were true and whose conclusion was false.**

We also improve our definition of "deductive logic" to:

> **Deductive logic is the science of the rules of truth-preserving transformations on statements.**

[10]**Record this definition in your soul.** It is *central* to understanding logic.

All that matters for logic is the relation between the truth of our premise-statements and the truth or falsity of our conclusion-statements. Every logic we shall consider will have this property.

A **deductive argument** can be:

1. *Valid*, with *true* premises and a *true* conclusion;
2. *Invalid*, with *true* premises and a *true* conclusion;
3. *Invalid*, with *true* premises and a *false* conclusion;
4. *Valid*, with *false* premises and a *true* conclusion;
5. *Invalid*, with *false* premises and a *true* conclusion;
6. *Valid*, with *false* premises and a *false* conclusion;
7. *Invalid*, with *false* premises and a *false* conclusion.

The one thing it *cannot* be (by the definition of "valid") is "*valid*, with *true* premises and a *false* conclusion". Notice that you can have an *invalid* argument whose premises and conclusion are all *true*, and *valid* arguments whose conclusions are *false*.

One more definition may be useful in light of the above. We call an argument (**not** a statement or belief) *sound* when it satisfies the definition:

A sound deductive argument is a valid argument whose premise(s) are true.

From the definitions, what can you say about the conclusion of a sound deductive argument?

1.2.4 Exercise on arguments

For each of the following informal arguments[11], identify the premises and the conclusion of the argument made. Write these in "standard form", meaning list the premises first, and the conclusion last, each statement on a separate line. Some statements will be neither (they will be intermediate parts of the argument from the premises to the conclusion); you should not include those in your answers.

1. It is right that men should value the soul rather than the body; for perfection of soul corrects the inferiority of the body, but physical strength without intelligence does nothing to improve the mind. (Democritus)
2. There cannot be any emptiness; for what is empty is nothing, and what is nothing cannot be. (Melissus)
3. About the gods, I am not able to know whether they exist or do not exist, nor what they are like in form; for the factors preventing my knowledge are many: the obscurity of the subject, and the shortness of human life. (Protagoras)
4. In the beginning man was born from creatures of a different kind; because other creatures are soon self-supporting, but man alone needs prolonged nursing. For this reason he would not have survived if this had been his original form. (Anaximenes)

[11]Taken from *The Logic Book* by Bergmann, Moor, and Nelson.

5. Let us reflect in another way, and we shall see that there is great reason to hope that death is good; for one of two things—either death is a state of nothingness and utter unconsciousness, or as men say, there is a change and migration of the soul from this world to another. Now if you suppose that there is no consciousness, but a sleep like the sleep of him who is undisturbed even by dreams, death will be an unspeakable gain ... for eternity is then only a single night. But if death is the journey to another place, and there, as men say, all the dead abide, what good, O my friends and judges, can be greater than this? ... Above all, I shall then be able to continue my search into true and false knowledge; as in this world, so also in the next; and I shall find out who is wise, and who pretends to be wise, and who is not. (Socrates)

1.3 Truth-Functional Connectives

Expressions used to link sentences to create a new *compound* sentence are called "**connectives**". "Not" is a connective, even though it is used with a single sentence rather than connecting two sentences.

Connectives actually link or modify *statements*, rather than *sentences*.[12] Many sentences, in many languages make the *same* statement. All that matters for logic is whether the conclusion of a deductive argument can be false when the premise(s) are true. We don't care about the particular words or language in which the premise-statement(s) and conclusion-statement are expressed. We don't care about the *sentences* used to make the *statements*.

The connectives themselves are not language-specific. The conjunction-connective can be expressed many ways, even in English (as we saw). "Et" and "und" and "y" and a whole bunch of words in other languages do the same logical job of conjoining two statements. Similarly "not" has equivalents in English ("He is not going" is equivalent to "It is not the case that he is going" and so on) and in other languages. We treat all connectives that do the same logical job as the same.

Some connectives are unimportant to logic.[13] We shall only investigate "truth-functional connectives", *viz* connectives with the following property.

By "**truth-functional connective**" we mean a connective which links statements or modifies a statement in such a way that the truth or falsity of the resulting compound statement (the original statement(s) plus the connective) is a function of (*i.e.*, depends only on) the truth or falsity of the (original) component statement(s).

A connective that is *not* truth-functional is the phrase "I believe that...". You can stick that phrase on the front of a sentence (*e.g.*, "Arwen is somewhat neat".) and get a new sentence ("I believe that Arwen is somewhat neat"). But the connective is *not truth-functional*. The truth or falsity of the new statement (made by the sentence "I believe that Arwen is somewhat neat") cannot be known just by knowing whether "Arwen is somewhat neat" is true or false. Other examples might include "... because ..." (for instance "John skipped class because Susan loves Jo"); the

[12] Although one should keep this distinction clear, there may be times when any of us might use "sentence" when really meaning "statement". You should be able to decide, based on the context, which is intended: is it the actual words used (the sentence), or is it their essence (the statement) that is at issue?

[13] Well, not exactly—what I really mean here is "unimportant to the classical logic we shall study". Many connectives that fall outside our survey are considered by other logics, for example modal logic, which studies non-truth-functional operators like "necessarily" and "possibly".

truth (or otherwise) of such a statement does not depend on the truth of the individual components. It is not truth-functional.[14]

On the other hand, the "not" connective *is* truth-functional. If it is true that Arwen is somewhat neat, then "Arwen is not somewhat neat" must be false.

Propositional logic is the logic of the truth-functional connectives. The laws of valid deductive reasoning in propositional logic are based on the meanings (rules for the correct use) of the truth-functional connectives, and on nothing else.

1.3.1 The symbolism of propositional logic

The formal sciences (logic and mathematics) develop their own languages. These artificial languages[15] are more precise and clear than natural languages. More precise and clearer language leads to more precise and clearer thought and concepts. Learning mathematics and modern logic involves learning its special language—the symbolism.

The main advantages of symbolic logic (using an artificial language of symbols) are: (1) we avoid writing a lot of stuff that doesn't matter for logic; (2) we emphasize the universality of logic; and (3) it makes it easier to recognize the *form* of a compound statement or of an argument.

The English sentence "Frodo is an airhead" makes a statement. Other sentences (in English or in other languages) make the *same* statement. We can refer to the statement they all make as "the airhead statement" or "what you said about Frodo". In propositional logic we use symbols, usually letters, to stand for particular simple statements. For example, the letter "A" can represent the statement made by the sentence "Frodo is an airhead". "B" could be symbolic shorthand for the statement made by "Arwen is somewhat neat".

Then we could use "A and B" as shorthand for the statement made by the compound sentence "Frodo is an airhead and Arwen is somewhat neat". "Not B" could stand for "It is not the case that Arwen is somewhat neat".

Often we refer to *any* statement or *a* statement without having any particular statement in mind. For example, we might want to say, "The negation of a true statement is false". When mathematicians want to say something about particular numbers they use symbols (called constants) like 12 or 12364. They use variable-symbols like x or y to stand for some number, or an unknown number, or any number. In propositional logic, we use letters like p and q to represent some (unspecified) statement, or any statement. We could symbolize a conjunction of two (unspecified) statements as "p and q". That way, we can talk about a conjunction without worrying about what statements are conjoined. We can discuss conjunctions in general, or the form of a conjunction. Such letters act as *statement variables*.

We call the truth or falsity of a statement its *truth-value*. Every statement can have one of two possible truth-values (true or false) and every meaningful statement has one of those values (even when we don't know what value it has).[16]

We define the truth-functional connectives in terms of the truth-value of the statement that results from using that connective with one or more statements.

[14] We shall see a variant of "because", namely "material implication", which is truth functional. The distinction comes from the fact that with material implication we drop all suggestion of causality, which is what stops "because" from being truth-functional.

[15] Called "artificial" to distinguish them from "natural" languages like English, French, *etc.*

[16] There are alternate logics which do in fact allow more truth values, but we shall not study them in this course.

1.3.2 The connectives

You must *understand* and memorize the four fundamental truth tables (for $\neg$, $\wedge$, $\vee$, and $\longrightarrow$) of this section. They form the basis of our approach to logic, and they define how these four basic logical operators behave.

Negation

The simplest truth-functional connective is the negation connective. We use the symbol "$\neg$" to stand for the word "not" and its logical equivalents. So we symbolize "Frodo is not an airhead" by $\neg A$.

Using our statement-variable symbols, we can define the negation operator as: "$\neg p$ is false when p is true, and $\neg p$ is true when p is false". The variable p tells us that this applies to any statement that you might substitute for p. So it tells us that $\neg A$ is false if A is true, and $\neg A$ is true if A is false, for any constant (representing a particular statement) A. It also tells us that $\neg B$ is false if B is true, and so on.

We **define** the $\neg$ connective with a **truth table**. The truth table for $\neg$ is:

p	$\neg p$
$\top$	$\bot$
$\bot$	$\top$

In a truth table, "$\top$" indicates the case where the statement-form p is replaced by a statement that is guaranteed to have the truth-value "true," (for instance the statement $1 = 1$), and "$\bot$" indicates the case where the statement-form p is replaced by a statement that is guaranteed to have the truth-value "false" (for instance the statement $1 = 0$). In other words, $\top$ is a generic true statement, and $\bot$ is a generic false statement. The truth table tells you exactly what the "$\neg$" connective *does* to the truth-value of any compound statement. **The whole logical meaning of "$\neg$" consists of what it does to the truth-value of a statement.** It is *truth-functional* because *it operates on a truth-value and produces a new truth-value according to a rule.* The rule is that it produces a true statement from a false one, and *vice versa.*

> A compound statement whose main connective[17] is the negation connective is a *negation.* The negation $\neg p$ is false when p is true and true when p is false.

Negation is a **unary** connective. It works with only *one* statement (which could be a compound statement). The following connectives are all **binary** connectives. Each works with two statements (which could be compound statements).

Conjunction

A conjunction is a compound statement of the form "p and q". Suppose you replace p with a true statement and replace q with a false statement. Is the conjunction of the true statement with a false statement true? A false conjunct makes the whole conjunction false. If *both* conjuncts are false, the conjunction is false. We would only say that a conjunction is true if **both** conjuncts were true.

> A compound statement that results from linking two (simple or compound) statements p and q with the conjunction connective $\wedge$ is a **conjunction**, and p and q are its

[17] The notion of a "main connective" will be developed in section 1.3.3, on formation rules.

conjuncts. A conjunction is true if and only if both conjuncts are true. It is false if and only if at least one conjunct is false.

We symbolize the conjunction connective with "$\wedge$".[18] The truth table defining $\wedge$ is:

p	q	$p \wedge q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\bot$
$\bot$	$\bot$	$\bot$

Since "$\wedge$" connects two statements, we have to define it *for all possible combinations of truth-values* of p and q. Start with the rightmost of the simple statement forms (q, in this case) and go down the column, alternating $\top$ and $\bot$. Then go down the column of the next simple form to the left (p), alternating *pairs* of $\top$s and $\bot$s. If there were three simple forms, the next column to the left would consist of four $\top$s followed by four $\bot$s, and so on. This technique ensures that every combination is included.

The English word "but" has exactly the same truth-functional meaning as "and". The two words are *truth-functionally equivalent*: if we construct a truth table for "and" and for "but", we would end up with the same table. That is, "Frodo is an airhead but Arwen is somewhat neat" is true or false in exactly the same circumstances as "Frodo is an airhead and Arwen is somewhat neat". The difference in meaning is not truth-functional, but reflects attitudes propositional logic does not attempt to capture. Symbolize "Frodo is an airhead but Arwen is somewhat neat" as $A \wedge B$, just as you would "Frodo is an airhead and Arwen is somewhat neat".

Disjunction

Another common English connective is "or". We symbolize this as "$\vee$".[19] It is truth-functional, but defining it presents a slight problem, since there are at least two meanings (one more common than the other) of the word "or".

Consider "Either you do all the assignments or you fail the course" ($D \vee F$). When is that statement false? Suppose you do all the assignments (D is true) and pass (don't fail) the course (F is false). You'd say the statement was true. Suppose you don't do all the assignments and you fail the course (D is false, but F is true). Still true. If you don't do all the assignments and don't fail the course (D is false and F is false), the statement is false. But what if you do all the assignments (D is true) and you fail anyway (F is also true)? Did the statement imply that both these eventualities could not occur? The answer is "yes and no, depending on what 'or' means".

The English word "or" is *ambiguous*: it has two (at least) possible meanings: **inclusive-or** and **exclusive-or**. "D or F" could mean "either D or F *or both*" (inclusive) or it could, more commonly, mean "either D or F *but not both*" (exclusive). When a highwayman holds up a stage and shouts "Your money or your life", the passengers hope that he intends to take their lives *or* their money *but not both*. But when a teacher says, "Do the work or flunk the course", he usually intends that either *or both* could happen: he is not *guaranteeing* a pass to everyone who works. (But notice something: if his statement is true, then he *is* guaranteeing a failure to anyone who doesn't do the work. In other words, if someone does not do the work *and* does not fail the course, then his statement was false. You will see this in the following truth table.)

[18]Different authors favour different symbols; & is not uncommon, since it's simple to find on a keyboard.
[19]This looks like the letter "v", but in a different typeface. It comes from the Latin "vel", meaning "or".

Such ambiguity is intolerable in logic. We have to decide which "or" we want to symbolize by "$\vee$". Logicians and mathematicians use the **inclusive** sense of "or", where it means "either or both". The truth table is, then:

p	q	$p \vee q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\top$
$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\bot$

> A compound statement whose main connective is $\vee$ ("vel") is called a **disjunction**, and its two component statements are called its **disjuncts**. A disjunction is true if and only if either or both of its disjuncts is/are true. It is false only if both disjuncts are false.

We could use another symbol (and truth table) for the exclusive sense of "or" ("exclusive-or") but we don't need it. "And" means the same (logically) as "but", so "either A or B but not both" is just $(A \vee B) \wedge \neg(A \wedge B)$. Later we'll use truth tables to show that this expression captures the sense of exclusive-or. It's an amusing exercise (see **BAFact 5** in Chapter 9) to define inclusive-or in terms of exclusive-or (and other connectives) in a similar fashion; primarily we use inclusive-or as the basic form not only because inclusive-or has very nice properties (*e.g.* it is dual to conjunction), but also because in mathematics (and science), we usually want the inclusive-or, and only rarely seem to need exclusive-or—one way in which mathematicians differ from highwaymen. It's important to remember this, since the inclusive-or is less common in daily non-mathematical usage.

Material implication

Consider the English **implication** or *conditional*[20] "If you marry me (then) I'll do all the cooking". When the **premise** (or *antecedent*) of the implication (the part between "if" and "then") is true (in this case, you do marry me) and the **conclusion** (or *consequent*) (the part after the "then") is false (I don't do all the cooking), we would say that the conditional statement is false. In the case where you marry me and I *do* do all the cooking, we would call it true.

What if the marriage does not take place? The premise of the implication is false. Should we say that the implication is true or that it is false?

It doesn't seem to matter whether the conclusion is true or false. In a sense "all bets are off!": you didn't marry me, so any promise I made on that basis no longer holds. I won't be breaking my word, regardless of what meals I do or don't cook. Our problem does not arise out of ignorance as to whether or not I do the cooking.

We might like to say that the implication is neither true nor false when the premise is false. But propositional logic requires that *every meaningful statement must have a truth-value*—must be either true or false, and not both. If the truth-value of an "if ... then ..." statement does not depend on the truth-values of its components (the premise and the conclusion), then "if ... then ..." is not a truth-functional connective. This would be a disaster for propositional logic, where implication is centrally important.

It would appear then that the ordinary-language "if ... then ..." conditional connective is (logically) ambiguous. We shall remove this ambiguity by *defining* "if ... then ..." truth-functionally. We call the resulting connective **material implication** or **the material conditional**, symbolized

[20] "Conditional" and "hypothetical" are nouns or adjectives logicians use for "if ... then ..." sentences. Often, especially in mathematical usage, we also call them "implications", and say "... implies ...".

using an arrow ($\longrightarrow$).[21]

The definition is:

p	q	$p \longrightarrow q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\top$

In the first two rows, this truth table confirms our natural tendencies about how to handle implication. The last two rows reflect the fact that we have *defined* the implication to be true whenever the premise is false. Generally, those last two rows correspond to situations where normal language probably doesn't consider the "if ... then ..." situation, so our definition doesn't interfere too badly with everyday usage (but more on that below!).

You will have noticed we use terminology for implications very reminiscent of the terminology for arguments. It is important to keep the following distinction clear: an implication or conditional is a **single** statement, with two components (think of one sentence with two subordinate clauses), but an argument is a collection of **several** statements. A premise of an argument is a complete statement, whereas the premise of an implication is only a sub-statement.

However, our terminology has the advantage of reflecting the similarity between a conditional statement and an argument. In a sense, a conditional statement collects the separate statements of an argument into a single statement with the same structure. Remember that an argument is valid if it is impossible (from the form of the argument) for a conclusion to be false if the premises are true. False premises do not make an argument invalid. This is similar to the structure of $\longrightarrow$ as given in the table above. There is a technical way to make this idea precise; you will find that in Remark 1.3.14 later.

You *must* remember the behaviour of "if ... then ..."; it is central to propositional logic, (just as is the related notion of a valid argument).

> The compound statement that results from linking two simple or compound statements with the "if ... then ..." connective $\longrightarrow$ is a **material implication** or *conditional*. The component statement that states the condition (in the "if ..." clause) is put before the $\longrightarrow$ and is called the **premise** or *antecedent*. The other component statement is put after the $\longrightarrow$ and is called the **conclusion** or *consequent*. The material implication is false if and only if the premise is true and the conclusion is false. It is therefore true in all other cases.

Logicians call the conclusion of a material implication the **necessary condition** for whatever the premise says, and they say that the premise is a **sufficient** condition for whatever the conclusion says. Think this statement over very carefully—it might seem counter-intuitive, but if $A \longrightarrow B$ is true, then the truth of A is sufficient to guarantee the truth of B, whereas the truth of B is necessary for A to be true (for if B were false and $A \longrightarrow B$ true, then A could not possibly be true).

Warning: In everyday English usage, there is a tendency to "misinterpret" implication, to assume that when one says "if A then B", what is meant is "A if and only if B", or in other words, that A and B are "equivalent", *i.e.* that they are either both true or both false. This is not the meaning of $A \longrightarrow B$ at all. Be very clear about this; it will probably require you to unlearn a meaning that seems very natural to you.

[21] This is a very important notion, and so it's perhaps no surprise that there are many different notations for it, including a double shafted arrow ($\Rightarrow$), a hook ($\supset$), a less-than symbol ($<$), and various other "weird arrows" ($\multimap$ is a favorite of mine), among others.

The implication $B \longrightarrow A$ is the **converse** of the implication $A \longrightarrow B$. The **contrapositive** of $A \longrightarrow B$ is $\neg B \longrightarrow \neg A$. The *converse* of a true implication could be false (or true), but the *contrapositive* of a true implication must be true. Notice that the contrapositive is just another way of saying the implication: $\neg B \longrightarrow \neg A$ has the same meaning, the same truth table, as $A \longrightarrow B$. These statements are "equivalent". (Check this for yourself!) By contrast, the converse is quite independent of the implication: $B \longrightarrow A$ and $A \longrightarrow B$ may both be true, both false, or one can be true and the other false, depending on the particulars of what A and B say. (Again, check this for yourself, with various examples for A and B.)

Remark: Is material implication paradoxical? Before leaving implication, let's consider the following seeming paradox. According to our definition of (material) implication, the following is a true statement: "If you are a purple unicorn, then you will win the lottery today". It is true simply because in fact, you are *not* a unicorn (purple or otherwise), and so whether or not you win the lottery has no bearing on the truth of this statement. Similarly, this is also a true statement: "If you are 50 years old, then $1+1=2$". This is true simply because $1+1=2$ is true (regardless of your age), and our definition of the (material) implication has the property that if the conclusion is true, the implication is true also, regardless of the truth of the premise. (Check it out: $\bot \longrightarrow q$ and $p \longrightarrow \top$ are always $\top$, regardless of what p and q are. Keep in mind *always* that the *only* way a implication is $\bot$ is $\top \longrightarrow \bot$.)

To most English-speakers, this seems ... well, silly! (though I'll say "paradoxical" to sound more impressive). Surely there ought to be some connection between the premise and the conclusion of an implication of this sort. But these examples play fast and loose with that expectation. And that's the nub of the matter: the paradox, if there is any, is merely one of *expectation.* We expect the English words "if ... then ..." to behave, in the formal setting of propositional logic, as they might in the informal setting of everyday English. *They do not.* It's as simple as that. The connectives $\neg$, $\wedge$, $\vee$, and $\longrightarrow$ are not words in everyday English, even if we pronounce them "not", "and", "or", and "if ... then ...". They are technical terms, whose meanings are strictly defined (by their truth tables), and on that there is no room for a difference of opinion. Sorry! If you think "or" should mean the exclusive variety, tough luck—that is not what $\vee$ means, and if it bugs you to call it "or", give it another name, like "vel". Similarly for $\longrightarrow$: call it "implies" or some other word you don't use very much. You cannot change the meanings of these connectives, however much you want to, without changing the very subject we are studying. If it makes you feel better, there are other logics which attempt to do just that, and you can turn to study them as a personal project *after* the semester is over! For example, "relevance logic" is a generic name for a family of logics which attempt to define an implication/conditional which requires some causal connection between antecedent and consequent. For now, however, we shall stay with classical propositional logic, with the connectives defined as above.

The biconditional—a derived connective

When an implication $p \longrightarrow q$ and its converse $q \longrightarrow p$ are both true, we say "p if and only if q". This kind of compound statement is important in mathematics. To represent this, we define another connective, the biconditional, written $\leftrightarrow$ or "$\equiv$" or even "iff", as in $p \leftrightarrow q$, in terms of the implication and its converse:

$$p \leftrightarrow q \ := \ (p \longrightarrow q) \wedge (q \longrightarrow p)$$

Whenever one sees $\equiv$ or $\leftrightarrow$, one should replace it as defined above.

In this definition, we had to use parentheses so that it was clear that the conjunction had implications as its two conjuncts. The ":=" symbol indicates that the expression on the right *defines* (*i.e.*, is the **definiens** for) the expression on the left (the **definiendum**). In logic, whenever we see the definiens, we can replace it with the definiendum, and *vice versa*. Since the two expressions have the same meaning, it is impossible for one statement to be false when the other is true. Replacing the definiens with the definiendum can never lead to an invalid (false conclusion from true premises) logical step. The truth table for $(p \longrightarrow q) \wedge (q \longrightarrow p)$ (exercise: construct it!—or look at section 1.3.9.) shows that this whole expression is true whenever p and q have the same truth-value (either both true or both false) and is false when p and q have different truth-values (one true and the other false).

1.3.3 Formation rules

The rules of the syntax of our symbolic language are the **formation rules**.

WFF is an abbreviation for *Well-Formed Formula*, often simply called a "formula". A simple statement (*e.g.*, A, or B, or Dusty is a cat) is a WFF. As we've defined the "$\neg$" connective, $\neg A$ is a WFF. But $A\neg$ is meaningless. It is not "well-formed". It is not a WFF. Neither is $A\neg C$. Neither is $\wedge A$, but $B \wedge A$ is a WFF. So is $A \wedge A$ (nothing in the definition of "$\wedge$" says that p and q have to be different statements). We shall adopt the convention that a statement variable is a WFF (since it obviously becomes a WFF whenever it is replaced by a simple statement or by a WFF).

The components of a compound statement may be compound. As we saw when we defined the biconditional, the conjuncts of a conjunction can be implications. But we may need parentheses to clarify the logical meaning of the resulting compound.

Is $\neg A \wedge B$ the conjunction of a negation and a simple statement or is it the negation of a conjunction? Is "$A \wedge B \vee C$" the conjunction of a simple statement A with a disjunction $B \vee C$ or the disjunction of a conjunction $A \wedge B$ and a simple statement C? We add **parentheses** and a **precedence rule** to the syntax of our symbolic language.

> **Parentheses Rule**: When a binary compound statement is used as a component of a compound statement, it must be surrounded by parentheses.

The example of conjoining (making a conjunction of) A and $B \vee C$ requires that we put parentheses around $B \vee C$ before using it as a conjunct. We get $A \wedge (B \vee C)$. Now it is clear that this is a conjunction and that its second conjunct is a disjunction.

> **Precedence rule**: the $\neg$ connective takes precedence over any other connective.

This means that $\neg A \wedge B$ must be interpreted as a conjunction whose first conjunct is a negation. It works like $(\neg A) \wedge B$. If we want to negate a conjunction, we must use parentheses to get $\neg(A \wedge B)$.

> **WFF rules**: A symbolic expression is a WFF if and only if:
>
> 1. it is a simple statement or a statement variable. Thus, A is a WFF. p is a WFF, $\top$ is a WFF, as is $\bot$. The statement represented by the sentence "It is raining" is a WFF. And so on. We call such WFFs "atomic formulas", "atomic propositions", or simply "atoms".
> 2. it is any WFF (in parentheses if it is a binary compound WFF) preceded by $\neg$. Any WFF formed according to rule (2) is a **negation**. Its **main connective** is the negation connective $\neg$.
> 3. it is any WFF (in parentheses if it is a binary compound WFF) followed by $\wedge$, followed by any WFF (in parentheses if it is a binary compound WFF). Any WFF formed according to rule (3) is a **conjunction** and its **main connective** is the

conjunction connective $\wedge$. The two WFFs linked by the connective are called **conjuncts**.

4. it is any WFF (in parentheses if it is a binary compound WFF) followed by $\vee$, followed by any WFF (in parentheses if it is a binary compound WFF). Any WFF formed according to rule (4) is a **disjunction** and its **main connective** is the disjunction connective $\vee$. The two WFFs linked by the connective are called **disjuncts**.
5. it is any WFF (in parentheses if it is a binary compound WFF) followed by $\longrightarrow$, followed by any WFF (in parentheses if it is a binary compound WFF). Any WFF formed according to rule (5) is an **implication** or conditional, and its **main connective** is the conditional connective $\longrightarrow$. The WFF before the connective is called the **premise** of the implication or the **sufficient condition**. The WFF after the connective is the **conclusion** of the implication, also called the **necessary condition**.

By rule (2), $\neg A$ is a WFF, because A is a WFF. Since $\neg B$ is a WFF by rule (2), so is $\neg\neg B$. So is $\neg\neg\neg\neg C$. So is $\neg(A \longrightarrow B)$ (since $A \longrightarrow B$ is a WFF according to rule (5)). Rule (3) says that $A \wedge B$ is a WFF. So is $A \wedge \neg B$. So is $\neg\neg B \wedge \neg(A \longrightarrow B)$. And so on. Rule (4) permits $A \vee B$ as a WFF. So is $A \vee \neg B$. So is $\neg\neg B \vee \neg(A \longrightarrow B)$. And so on. By Rule (5), $A \longrightarrow B$ is a WFF. So is $A \longrightarrow \neg B$. So is $\neg\neg B \longrightarrow \neg(A \longrightarrow B)$. And so on.

There is a second precedence rule we shall frequently use:

> **Second Precedence Rule**: $\longrightarrow$ binds less strongly than $\wedge$ and $\vee$ (which bind equally strongly as each other).

So, we can unambiguously interpret $A \longrightarrow B \wedge C$ to mean $A \longrightarrow (B \wedge C)$, and not $(A \longrightarrow B) \wedge C$, which must be bracketed as shown.

Remark concerning parentheses. Although one needs to be careful about parentheses, one has also to keep a sense of proportion about them. Their purpose is simply to avoid ambiguity in logical expressions, nothing more. We shall feel free to use different styles of bracketing when it helps make the expressions clearer. For example, contrast the following two expressions (intended to be regarded as identical):

$$(((A \wedge p) \longrightarrow (q \vee B)) \longrightarrow ((\neg A \wedge q) \vee \neg(B \longrightarrow \neg p))) \longrightarrow r$$

$$([(A \wedge p) \longrightarrow (q \vee B)] \longrightarrow [(\neg A \wedge q) \vee \neg(B \longrightarrow \neg p)]) \longrightarrow r$$

The use of brackets [ ] instead of parentheses ( ) helps keep track of what groups go together.

In fact, parentheses may be totally avoided if we adopt a different technical presentation (known as "reverse Polish notation"), where the connective *follows* the two formulas it joins, as in $AB\wedge$. With this presentation, the formula above would look like this:

$$Ap \wedge qB\vee \longrightarrow A\neg q \wedge Bp\neg \longrightarrow \neg\vee \longrightarrow r \longrightarrow$$

but for most folks, the current ("in-fix", *i.e.* connectives in between their arguments) presentation is clearer, even if it means fiddling with parentheses.

If we ignore parentheses for the moment, we can summarize the WFF rules in the following compact form:

$$P := A \mid \neg P \mid P \wedge P \mid P \vee P \mid P \longrightarrow P$$

meaning a proposition (WFF) is either an atomic proposition (a constant or a variable), the negation of a proposition, the conjunction of two propositions, the disjunction of two propositions, or the implication of two propositions.

1.3.4 Parsing complex symbolic expressions

"Parsing" is the process of analyzing an expression to discover (1) if it is well-formed (syntactically correct, grammatical) and (2) supposing it is well-formed, what kind of statement it symbolizes (is it a negation, a conjunction, a disjunction, or an implication—in other words, what is its main connective?).

To parse a compound expression, start by counting the left parentheses and the right parentheses. If there is not the same number of each, and if they are not balanced, the expression is not a WFF.

You probably remember "balanced parentheses" from high-school maths, but if not, here is a simple test you can use to check for them. Start at the left end of the expression, and count parentheses, beginning with 1 and adding "+1" for each "(" and "−1" for each ")". Here's an example, with the counting numbers indicated as superscripts:

$$(^1(^2(^3A \wedge p)^2 \longrightarrow (^3q \vee B)^2)^1 \longrightarrow (^2(^3\neg A \wedge q)^2 \vee \neg(^3B \longrightarrow \neg p)^2)^1)^0 \longrightarrow r$$

You should never get a negative number, and you should end up with 0, if the parentheses are balanced. Note that each matched pair of parentheses carry indices n, $n-1$.

If the expression passes the parentheses-check, proceed as follows:

Every simple statement symbol or statement variable in the expression is a WFF, by Rule (1). Underline or highlight the simple statement symbols and statement variables. Then follow the algorithm (where "highlight" means "highlight or underline"):

1. If the only thing that is not highlighted is a single connective, go to step 5.
2. If everything enclosed by a pair of parentheses is highlighted, highlight the parentheses and go to step 1. Otherwise, go to step 3.
3. If any highlighted component is immediately preceded by a ¬, extend the highlight to include the ¬ and go to step 1. Otherwise, go to step 4.
4. If there are two highlighted components separated by a binary connective (either ∧, ∨, or ⟶), extend the highlighting on the components to include the connective and go to step 1. Otherwise, go to step 2.
5. The connective that is not highlighted is the main connective. If the main connective is ¬, the expression is a negation; if ∧, it is a conjunction; if ∨, the expression is a disjunction; if ⟶, it is an implication.

Suppose we see an expression like

$$(D \vee \neg H) \longrightarrow (R \wedge (S \vee T)).$$

Here there are three left and three right parentheses and they are balanced, so this expression passes the parentheses-test. Highlighting the simple statement symbols, we get

$$(\underline{D} \vee \neg \underline{H}) \longrightarrow (\underline{R} \wedge (\underline{S} \vee \underline{T}))$$

There are several connectives that are not highlighted, so we go on to step 2. No highlighted thing has parentheses on both sides of it, so go to step 3. In step 3 of the algorithm, we get

$$(\underline{D} \vee \underline{\neg H}) \longrightarrow (\underline{R} \wedge (\underline{S} \vee \underline{T})).$$

We go to step 1. Several connectives are not highlighted. Nothing happens in steps 2 or 3. Step 4 gives

$$(\underline{D \vee \neg H}) \to (\underline{R} \wedge (\underline{S \vee T}))$$

(Notice that we did not highlight the $\wedge$ connective between R and $(S \vee T)$ because there was more than just a connective between the highlighted R and the highlighted $S \vee T$. There was also a left-parenthesis.) Back to step 1. There are still connectives that are not highlighted, so step 2 gives

$$\underline{(D \vee \neg H)} \to (\underline{R} \wedge \underline{(S \vee T)}).$$

Back to step 1. There are two connectives that are not highlighted, so we go to steps 2 and 3. There is no $\neg$ immediately preceding a highlighted component, so we go to 4. In $\underline{R} \wedge \underline{(S \vee T)}$ we have two highlighted components separated by a binary connective, so we extend the highlighting, getting

$$\underline{(D \vee \neg H)} \to (\underline{R \wedge (S \vee T)}).$$

One connective and two parentheses are not highlighted, so we go to step 1, then 2, and extend the highlighting on the right side to include the parentheses and back to 1. The highlighted expression now looks like

$$\underline{(D \vee \neg H)} \to \underline{(R \wedge (S \vee T))}.$$

The only thing that is not highlighted is the $\to$. We go to step 5. $\to$ is the main connective. The expression is an implication. Its premise is $D \vee \neg H$, which is a disjunction (parse it and see). Its conclusion is $R \wedge (S \vee T)$, a conjunction whose second conjunct is a disjunction.

OK—take a breath!! You will find with a little practice that in fact this takes about 5 seconds to do, really. Scan the displays above, noting how the underlines grow out from the simple statements to engulf more and more of the structure, by identifying the role each bit plays in the whole expression, ending up with two underlined bits joined by the main connective. In a way, every time a connective is put in parentheses it is "buried" lower in the structure, and the "topmost" connective is the main connective. It's usually pretty obvious which one that is—just match the parentheses and check that each connective joins exactly two (one for $\neg$) expressions.

1.3.5 Examples

1. Parse $\neg(C \wedge D) \vee (A \vee M)$. Two left and two right parentheses: check. Start with $\neg(\underline{C} \wedge \underline{D}) \vee (\underline{A} \vee \underline{M})$. Go to step 2, then 3: there is no highlighted component preceded by $\neg$ (there is an un-highlighted parenthesis after the $\neg$). Step 4: $\neg(\underline{C \wedge D}) \vee (\underline{A \vee M})$. Step 2 gives: $\neg\underline{(C \wedge D)} \vee \underline{(A \vee M)}$. From step 3: $\underline{\neg(C \wedge D)} \vee \underline{(A \vee M)}$. There is now only one un-marked connective. The expression is a disjunction. The left disjunct is a negation of a conjunction. The right disjunct is a disjunction.

2. Parse $\neg(\neg B \wedge \neg A)$. Following the algorithm, we get: $\neg(\neg\underline{B} \wedge \neg\underline{A})$, leading to $\neg(\underline{\neg B} \wedge \underline{\neg A})$, then $\neg(\underline{\neg B \wedge \neg A})$, and finally $\neg\underline{(\neg B \wedge \neg A)}$. The expression is a negation (of a conjunction, each of whose conjuncts is a negation).

1.3.6 Parsing and WFF exercise

1. Each of the following expressions is a WFF. Parse it and say what kind of WFF (negation, conjunction, disjunction, or implication) it is:
   (a) $\neg(A \vee \neg A)$
   (b) $A \wedge \neg(B \vee C)$
   (c) $\neg(A \to B) \to ((A \wedge \neg B) \vee \neg A)$
   (d) $\neg(A \wedge B) \vee \neg(\neg C \to \neg D)$

2. Make the longest WFF you can using only the symbols given (plus any number of parentheses you need). You don't have to use all the symbols shown, but you cannot use any symbol more often than explicitly shown. (For example, in (a), you can use two As, one B, one $\wedge$, one $\vee$, and two $\neg$s.) Check (parse) each WFF you construct and say what kind of WFF it is.
   (a) $A\ A\ B\ \wedge\ \vee\ \neg\ \neg$ (b) $A\ A\ \neg$ (c) $G\ H\ W\ N\ \neg\ \neg\ \vee\ \wedge\ \longrightarrow$
   (d) $D\ E\ F\ \neg\ \longrightarrow$ (e) $A\ B\ C\ \longrightarrow\ \longrightarrow\ \vee$ (f) Make up some of your own.

3. *Remark*: Technically, any expression containing a subexpression of the form $p \leftrightarrow q$ is not a WFF, since one must expand the subexpression according to the definition of $\leftrightarrow$. However, it is easy to show that we could add rules for $\leftrightarrow$ to the WFF formation and parsing rules, analogous to the rules for $\longrightarrow$, which would allow us to treat $\leftrightarrow$ as if it were a connective in the usual way. With such new rules, any expression would be a WFF if and only if the expression with the $\leftrightarrow$ expanded (using its definition) is a WFF, so nothing is a WFF with the new $\leftrightarrow$ rule that shouldn't be a WFF. Verify this claim. (As an exercise, this is optional, but you should understand the result.)

1.3.7 Substitution instances

> **A particular statement S is a substitution instance of a statement form F if S is the result of replacing every simple statement variable in F with a *simple or compound* statement constant.** None of the connectives in F may be altered or eliminated. If any statement variable occurs more than once in F, every occurrence must be replaced by the same statement constant in S.

We can replace a *simple* variable with a *compound* constant. $\neg((A \wedge B) \longrightarrow C)$ is a substitution instance of $\neg q$, because it results from replacing every distinct simple statement variable (*i.e.*, q) in $\neg q$ with the compound statement constant $(A \wedge B) \longrightarrow C$. The same statement $\neg((A \wedge B) \longrightarrow C)$ is also a substitution instance of $\neg(q \longrightarrow r)$, replacing the statement variable q with the constant $A \wedge B$ and the variable r with the constant C (and following the parentheses rule). So a single statement may be a substitution instance of many statement forms. It is also true that many statements may be substitution instances of the same statement form; for instance, $\neg(A \longrightarrow C)$ is another substitution instance of $\neg q$.

Note 1: The definition of "substitution instance" permits us to obtain a substitution instance from a statement form by replacing two *different* statement variables with the *same* statement constant. $A \vee A$ is a substitution instance of $p \vee q$.

Note 2: The definition does *not* permit replacing a compound statement form with a simple statement constant. $A \wedge B$ is not a substitution instance of $(p \vee q) \wedge r$, because A is not a substitution instance of $p \vee q$. The $\vee$ connective and the third distinct statement variable are lost in that substitution.

Note 3: The definition requires that two simple forms that are represented by the same variable be replaced by the same statement constants, so $A \vee \neg B$ is not a substitution instance of $p \vee \neg p$ (one p is replaced with A and the other with B).

Note 4: The definition of substitution instance is strictly *syntactic* (it depends on exactly what symbols are used and how they appear), not *semantic* (it does not in any way involve equivalence of statements). So, even though $\neg(A \vee B)$ is equivalent to $\neg A \wedge \neg B$, *i.e.* even though they always have the same truth value, the first is a substitution instance of $\neg p$, but the second is not, and the second is a substitution instance of $p \wedge q$, but the first is not. The notion of substitution instance is not a matter of truth values, but rather one of the actual symbols used.

Note 5: You should notice the role of parsing here: a statement S can only be a substitution instance of a statement form F if both S and F have the same main connective, and this must hold *recursively* as you go "deeper" into the structure of F, so for example, if both S and F are conjunctions, then the first and second conjuncts of each must also have the same main connective. This only ceases to be a condition when you reach variables in F. I won't make this statement too technical: look at the examples and you should see what is meant here. (This is really the main reason we spent time on parsing: recognising substitution instances is crucial for the next chapter, but we'll never look at parsing for its own sake again.)

Any simple statement is a substitution instance of the simple statement form p. So is any compound statement, like $(W \rightarrow A) \leftrightarrow ((A \wedge B) \vee W)$ (it results from replacing p with that whole long thing). And so is $\neg\top$. But A is *not* a substitution instance of $\neg p$ (we lost the negation connective). $(A \vee B) \wedge (W \rightarrow \neg\bot)$ is a substitution instance of $p \wedge q$ because (as parsing shows) it is a conjunction, and $p \wedge q$ is the form of a conjunction. $A \wedge \neg A$ is also a substitution instance of $p \wedge q$. So is $H \wedge H$. $(W \wedge X) \rightarrow J$ is not (it's an implication, and the form $p \wedge q$ is the form of a conjunction).

This may seem complicated and unfamiliar, but it is really quite simple, once you get the idea—it should help if you practice the exercises. Identifying substitution instances is crucial to the ability to show that an argument is valid when it is represented in the symbols of truth-functional logic.

1.3.8 Exercise on statement forms and substitution instances

For each statement form in the left-hand column, say which of the statement constants in the right-hand column are substitution instances of that form.

a. p	1. A
b. q	2. $A \rightarrow B$
c. $\neg p$	3. $(A \vee B) \rightarrow C$
d. $p \rightarrow q$	4. $(\neg A \vee B) \rightarrow C$
e. $\neg p \rightarrow q$	5. $\neg(A \vee B) \rightarrow C$
f. $\neg(p \rightarrow q)$	6. $\neg(\neg A \vee B) \rightarrow C$
g. $\neg(\neg p \rightarrow q)$	7. $\neg((A \vee B) \rightarrow C)$
h. $(p \vee q) \rightarrow r$	8. $\neg(\neg(A \vee B) \rightarrow C)$
i. $(p \vee p) \rightarrow \neg r$	9. $\neg(\neg(\neg A \vee B) \rightarrow C)$
j. $(\neg p \vee q) \rightarrow r$	10. $\neg((\neg A \vee B) \rightarrow C)$
k. $\neg(\neg p \vee q) \rightarrow r$	
l. $\neg(p \vee q) \rightarrow r$	
m. $\neg(\neg(\neg p \vee q) \rightarrow r)$	
n. $\neg((p \vee q) \rightarrow r)$	
o. $\neg(\neg(p \vee q) \rightarrow r)$	

1.3.9 Truth tables

To make a truth table of a compound statement type, we must look at the form of that type of statement and construct a truth table that applies to every statement of that form. For this reason, we only use statement variables in the WFFs for which we construct truth tables. (We allow one exception: we may usefully construct truth tables for WFFs which contain the constants $\top$ or $\bot$, and as many variables as we wish: for example, we might construct the truth table for the WFF $(p \vee \bot) \rightarrow (\bot \wedge \neg q)$. Why not do so now as an exercise?)

Earlier we saw that we don't need an exclusive-or connective, because $(p \lor q) \land \neg(p \land q)$ did the same truth-functional job (and therefore has the same meaning) as "p exclusive-or q". This is illustrated by its truth table. (We shall explain the meaning of the various columns in a moment.)

p	q	$(p \lor q)$	$\land$	$\neg$	$(p \land q)$
$\top$	$\top$	$\top$	$\bot$	$\bot$	$\top$
$\top$	$\bot$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
			*		

To construct this truth table, we start by making columns for all the simple statement-forms we'll need. The truth-values for the right-most simple form q are assigned alternating $\top$s and $\bot$s. The next column to the left gets alternating pairs of $\top$s and $\bot$s. If there is a third column, it gets four $\top$s followed by four $\bot$s, and so on. Then we do the columns for the simplest compound statement-forms—the next ones that we underline or highlight when parsing. We look at the truth table for the connective in the compound form (*e.g.*, in the example above we look first at the truth table for disjunction). That truth table determines what we put under the $p \lor q$, based on the values of the components p and q. We then do the column for the next-simplest component (the simple conjunction), and then the column for its negation. Finally we do the column for the most complex compound statement-form. The column $\bot$ $\top$ $\top$ $\bot$ under this last form (the column that contains the main connective of the whole statement) is called the "**main column**" of the truth table. We have put an asterisk under the main column to draw your attention to it. We use the truth table for the main connective ($\land$) to decide whether to put a $\top$ or a $\bot$ into the main column.

We defined the biconditional connective as the conjunction of a conditional and its converse: any expression of the form $p \leftrightarrow q$ is defined to mean the same as an expression of the form $(p \longrightarrow q) \land (q \longrightarrow p)$. The truth table for the biconditional then is:

p	q	$(p \longrightarrow q)$	$\land$	$(q \longrightarrow p)$
$\top$	$\top$	$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$	$\bot$	$\top$
$\bot$	$\top$	$\top$	$\bot$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

To enter the values for the $q \longrightarrow p$ column, remember that an implication is only false when its premise is true and its conclusion is false. Look for cases where q has value $\top$ and p has value $\bot$ in the first two columns. We find it only on the third row, so we enter $\bot$ in the third row of the column, and put $\top$ on the other rows. The main connective is $\land$, which is only true when both conjuncts are true. We look for the cases where both $p \longrightarrow q$ and $q \longrightarrow p$ are true (the first and last rows) and put a $\top$ there, and put $\bot$ on the other rows. From this truth table we see that the truth table for $p \leftrightarrow q$ is:

p	q	$p \leftrightarrow q$
$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\bot$
$\bot$	$\bot$	$\top$

As a shortcut, we may now use this truth table whenever we meet the biconditional in a WFF, instead of expanding the biconditional *via* its definition.

Looking at these tables, we can notice some properties of the WFF forms they represent. For instance, we see that the main column of the exclusive-or truth table $(p \vee q) \wedge \neg(p \wedge q)$ is exactly the opposite of the main column for the biconditional. That means that we could get the exclusive-or result either by a WFF of the form $(p \vee q) \wedge \neg(p \wedge q)$ or by a WFF of the form $\neg(p \leftrightarrow q)$. This also tells us that we could have defined the biconditional as $\neg((p \vee q) \wedge \neg(p \wedge q))$ (the negation of exclusive-or). (In the exercises below you will see this is also equivalent to $(\neg p \wedge \neg q) \vee (p \wedge q)$; in words: "$p$ is equivalent to q iff either both p and q are false, or they are both true". The truth table makes this very clear.)

Caution: Looking at a WFF like $\neg(A \vee B)$, some students do what you would do with an arithmetic expression like $-(a + b)$, which is to "multiply through by minus one". But $\neg$ is not a minus sign, and this kind of move would be a mistake. Exercises 4, 5 below let you look at the truth tables for the forms of similar-looking expressions that have different truth tables (and so, different meanings). There is an algebra of truth values, which we shall see in Chapter 9, at the end of the semester, but although it has some similarities with ordinary high-school algebra, there are very important differences as well, and you should not confuse them.

1.3.10 Exercise on truth tables

1. Construct the truth table for $\neg((p \vee q) \wedge (q \longrightarrow \neg p))$.

2. Construct the truth table for $(p \wedge q) \vee (\neg p \wedge \neg q)$.

3. Compare the main columns of the truth tables you constructed in 1 and 2 with the main column for the biconditional truth table. What do you observe? Since a truth table defines the meaning of an expression, these three expressions have the same meaning for logic. We say that they are equivalent (see the next section).

4. Compare the truth tables for $\neg p \wedge q$, $\neg(p \wedge q)$, $\neg p \wedge \neg q$ and $\neg p \vee \neg q$. Draw conclusions about the similarities or differences between them.

5. Compare the truth tables for $p \wedge (q \vee r)$ and $(p \wedge q) \vee (p \wedge r)$; do the same for $p \vee (q \wedge r)$ and $(p \vee q) \wedge (p \vee r)$. You should find that each pair has the same truth table, so we have two more equivalences. The first equivalence pair looks rather like the high-school algebra distributive law: $a \times (b + c) = (a \times b) + (a \times c)$. However, under this analogy, the second equivalence would seem to say $a + (b \times c) = (a + b) \times (a + c)$, which certainly is not true of high-school algebra. (Try it with numbers: *e.g.* is $1 + (2 \times 3) = (1 + 2) \times (1 + 3)$?) As we said above, the algebra of propositions is quite different from the algebra of numbers, in spite of some similarities. Don't try to use your high-school algebra here!

1.3.11 Truth-functional equivalence

When the main columns of the truth tables for two expressions are the same, we say that the two expressions are **truth-functionally equivalent**. Using one expression rather than the other makes no difference in truth-functional logic.

Remark: In view of the nature of the truth table for $\leftrightarrow$, it should be clear that two expressions P and Q are truth-functionally equivalent (or simply "equivalent") if the truth table for $P \leftrightarrow Q$ has the property that its main column has only $\top$ as truth value. (We call such an expression a "tautology", as you will see soon.) Informally, you may remember this as "P and Q are equivalent if each implies the other".

Sometimes when you translate a statement from a natural language into the language of the symbolism, you may discover two (or more) alternative translations. Which is correct? A truth table will show whether the different-seeming translations are equivalent. If they are, the best translation is the one that best captures the "feel" of the original statement. If not, choose the translation that best captures the truth-functional meaning of the natural-language statement.

We can symbolize the same natural-language compound statement using different connectives. We could have used fewer connectives than the four (plus the defined biconditional connective) that we have. Just one carefully chosen connective can be used to do everything that we do with our five. It is more difficult to translate from natural language to the symbolism when we use fewer connectives. Our set of connectives is a compromise between simplicity (why we left out the exclusive-or connective) and ease of translation.

1.3.12 Some equivalences

Here is a list of useful equivalences; try constructing the truth tables for some of these to verify that they are indeed equivalences. Think of the "intuitive meaning" for each statement, and try to see why it must be an equivalence.

1. Commutativity:
   (a) $(p \wedge q) \leftrightarrow (q \wedge p)$ (b) $(p \vee q) \leftrightarrow (q \vee p)$
2. Associativity:
   (a) $((p \wedge q) \wedge r) \leftrightarrow (p \wedge (q \wedge r))$ (b) $((p \vee q) \vee r) \leftrightarrow (p \vee (q \vee r))$
3. Distributivity:
   (a) $((p \wedge q) \vee r) \leftrightarrow ((p \vee r) \wedge (q \vee r))$ (b) $((p \vee q) \wedge r) \leftrightarrow ((p \wedge r) \vee (q \wedge r))$
4. De Morgan Laws:
   (a) $\neg(p \wedge q) \leftrightarrow (\neg p \vee \neg q)$ (b) $\neg(p \vee q) \leftrightarrow (\neg p \wedge \neg q)$
5. Others:
   (a) $(p \longrightarrow q) \leftrightarrow (\neg p \vee q)$ (b) $\neg(p \longrightarrow q) \leftrightarrow (p \wedge \neg q)$

1.3.13 Tautologies, contradictions, and contingencies

The truth table for $p \wedge \neg p$ is

p	p	$\wedge$	$\neg p$
$\top$		$\bot$	$\bot$
$\bot$		$\bot$	$\top$

which shows that a statement of the form $p \wedge \neg p$ is false whether p represents a true or a false statement.

> A *statement-form* (made of variables and connectives) is a **contradiction** if and only if its truth-value is $\bot$, no matter what the truth-values of its component statements. A *particular statement* is a contradiction (it is **contradictory**, or self-contradictory) if it is a substitution instance of a contradictory form.

So any statement that is a substitution instance of the form $p \wedge \neg p$ is a contradiction. Contradictions are **trivially false.** "Trivial" means that their truth doesn't depend on the truth of any particular statements, and therefore it doesn't depend on the way the world is.

Now look at $(p \wedge q) \to p$.

p	q	$(p \wedge q)$	$\to$ p
$\top$	$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$	$\top$
$\bot$	$\top$	$\bot$	$\top$
$\bot$	$\bot$	$\bot$	$\top$

Statements of the form $(p \wedge q) \to p$ are true, no matter what truth-values p and q have.

> A *statement-form* is a **tautology** (is tautologous) if and only if its truth-value is always $\top$. By extension, a *particular statement* is a tautology if it is a substitution instance of a tautologous statement-form.

Tautologies are **trivially true**. $(p \wedge q) \to p$ is a tautology (it is **tautologous**). A common example of a tautologous statement-form is $p \vee \neg p$, whose truth table is:

p	p	$\vee$	$\neg p$
$\top$		$\top$	$\bot$
$\bot$		$\top$	$\top$

Finally, some (indeed, most) statement-forms have neither of these properties, and take on both $\top$ and $\bot$ as truth values, depending on the truth values of the constituent parts. We call these contingencies.

> A statement-form is a **contingency** if and only if it is neither a tautology nor a contradiction. We call a particular statement **contingent** when it is not a substitution instance of any tautologous or contradictory form. Such statements are **non-trivial** (although they may not be important).

One way to show that a statement is a tautology or a contradiction is to construct a truth table of its statement form. We re-write the statement using variables, if necessary, replacing each distinct simple statement constant (other than $\top$ and $\bot$) with a distinct simple variable, and using the same simple variable for every occurrence of the same simple constant. The statement form should exactly duplicate all connectives and parentheses that are in the original compound statement.

For example, given the statement $((A \to M) \wedge (M \to L)) \to (A \to L)$, we could either treat A, M, L as variables, or we could construct the truth table for the form $((p \to q) \wedge (q \to r)) \to (p \to r)$. The form is exactly like the statement we're checking, except every A in the statement is replaced with a p, every M is replaced with a q, and every L is replaced with an r.

The statement form contains three simple variables, p, q and r. We start with a column for each of these. Beginning with the rightmost column (r), we alternate $\top$ and $\bot$ values. Moving to the next column to the left (q), we write alternating pairs of $\top$s and $\bot$s. Moving one more column to the left (p), we write alternating sets of four $\top$s and $\bot$s. This mechanical procedure ensures that our truth table will contain every possible combination of truth-values for the three variables. To the right of those three columns we make a column for each compound form that is a component of the whole compound form: this gives us columns for $p \to q$, for $q \to r$, and $p \to r$, then finally for $(p \to q) \wedge (q \to r)$, and the main connective, *i.e.* for the whole expression. (Since this is our first example of a truth table with three variables, we have indicated the order in which we fill the columns with little numbers above them; as before, we indicate the main column with an asterisk.)

			(1)	(4)	(2)	(5)	(3)
p	q	r	$((p \to q)$	$\wedge$	$(q \to r))$	$\to$	$(p \to r)$
$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\top$	$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$	$\bot$
$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\bot$	$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$
						*	

Every row of the main column contains a $\top$. That means that an expression of the form $((p \to q) \wedge (q \to r)) \to (p \to r)$ is true no matter what the truth-values of its component simple parts (p, q and r). It is a tautologous form. Any statement that is a substitution instance of this form is a tautology. Our original statement is a substitution instance of this form, so it is a tautology.

1.3.14 Remark

Earlier we remarked on the similarity between material implications and argument-forms. For example, an argument of the form “Premise: P; Conclusion: C” is valid if and only if the statement $P \to C$ is a tautology. If the argument has several premises: “Premise: P_1; Premise: P_2; Premise: P_3; Conclusion: C” for instance, then we can “internalize” this (represent it as a single statement) by $(P_1 \wedge P_2 \wedge P_3) \to C$. The argument is valid if and only if the corresponding conditional sentence is a tautology. In general, the **validity** of any argument-form can be expressed as whether or not the corresponding conditional statement is tautological. It is in this sense that material implication “internalizes” valid argument.

As an example, *via* the previous truth table we have just established a form of valid argument, namely if one has premises of the form $p \to q$ and $q \to r$, and a conclusion of the form $p \to r$, then the argument is valid. Every tautology corresponds to a valid form of argument in such a manner.

The method of showing that a statement is a contradiction is exactly parallel. If a truth table constructed according to the method above results in a main column that contains both $\top$s and $\bot$s, the statement is contingent.

1.3.15 Exercise on tautology, contradiction and contingency

For each of the following statements, determine (show and say) whether it is a contingency, a tautology, or a contradiction. In each case write an English interpretation of the statement.

1. $(A \to B) \to \neg A$
2. $(A \wedge B) \wedge \neg A$
3. $A \wedge \neg(B \vee A)$
4. $(A \to B) \leftrightarrow (\neg B \to \neg A)$
5. $(A \to B) \to (B \vee \neg A)$
6. $(A \to B) \vee (B \to A)$

1.4 Translation

Translating from one natural language (like English) to another (*e.g.*, French) is difficult. You have to be familiar with both languages and with all sorts of subtle nuances of expression and idiom. Artificial languages like symbolic logic are much simpler than natural languages. Logical symbolism is less ambiguous or vague than a natural language. This greater precision can be a problem when translating from informal natural languages into the formal language of the symbolism.

Some people think that this shows that the symbolism is an inadequate language. They are wrong. It is the natural languages that are inadequate—for logic. Natural languages are far more powerful and expressive than any artificial language. But they were invented for use in a wider range of human activities than just doing logic. For the restricted uses of logic, natural language is too vague and ambiguous.

The problem is that logic[22] needs to restrict itself to the truth-functional aspect of language. Many natural-language expressions do several jobs over and above the truth-functional-connecting job. The non-truth-functional components of the meanings[23] of these expressions obscure the truth-functional meaning. It can be difficult to dig out just the truth-functional part.

For example, look again at the words "and" and "but". Truth-functionally, they have exactly the same meaning. For logic, there is no difference between "Rosemary is an attractive woman and she's a lawyer" and "Rosemary is an attractive woman but she's a lawyer". The two sentences have somewhat different meanings, but there is no difference for logic. The difference is not truth-functional.

In natural languages, words that sometimes do one (truth-functional) job may also play another role in the language. For example, "and" may be used to connect two sentences in a conjunction-sentence, or it may be used to conjoin two expressions to make the compound *subject* (or object) of a *simple* sentence. "Monica and Steffi are tennis players" makes a compound statement, whose logical meaning is "Monica is a tennis player and Steffi is a tennis player". The "and" is the truth-functional conjunction connective we symbolize with "$\wedge$". But the sentence "Monica and Steffi are rivals" does *not* make the compound statement "Monica is a rival and Steffi is a rival". It makes a *simple* statement about a relation between two people—"Monica is a rival of Steffi".

Natural languages require subtle changes in the way a statement is expressed for *grammatical* reasons that have nothing to do with the *logical* meaning of the sentence. The result may be that two sentences that appear quite different may both have the same logical meaning (the differences being merely grammatical or stylistic).

Sometimes translation is difficult because the statement of an argument in a natural language may include one or all of the premises and the conclusion in a single sentence. In such cases, you have to figure out whether the conclusion is (1) the whole single compound statement or (2) a part of the sentence, so that other parts will be translated as one or more premise-statements.

When translating from natural language into the symbolism of propositional logic, the trick is to check and re-check that the truth-value of the translation behaves exactly like the truth-value of the form of the original sentence. If the translated statement would be true or false in exactly the same circumstances as the original, the translation has the same logical meaning as the original.

An additional requirement for good translation is that the symbolic representation should reflect as closely as possible the simplicity or complexity of the original sentence. For example, "If you love me and I'm able, I'll return to you" makes the same statement as "If you love me then, if I'm able, I'll return to you". The first could be symbolized as $(L \wedge A) \rightarrow R$ and the second as $L \rightarrow (A$

[22]At least, the kind of logic that we study in this course.

[23]By "meaning" in this text, I generally mean "use" or "rule for the correct use" of some piece of language.

$\rightarrow R$). These have identical logical meaning (their forms have the same truth tables—check it!). But the first translation more closely reflects the logical "feel" of the English sentence.

Once an argument has been represented in the symbolism, the problems of natural language (for logic) disappear. Translating a natural-language sentence or argument into the symbolism clarifies the logic of the sentence or argument. It is then easier to construct good arguments and to recognize bad (invalid) arguments. The symbolism is a more logical language than any natural language. It is pathetically poor for writing love songs.

Warning: There is one thing that might confuse you in making translations. You will know that "if p then q" will become $p \rightarrow q$. Other statements that also will become $p \rightarrow q$ include "q if p", "p implies q", and (likely only in a mathematical context) "p is a sufficient condition for q". But what about an expression of the form "p only if q"? Be careful: this is *not* $q \rightarrow p$, but is instead $p \rightarrow q$. This is one place where the statement following the "if" is *not* the premise, but instead is the conclusion. Another way to say this is that "q is a necessary condition for p" is also translated $p \rightarrow q$. (We did discuss this when introducing the material implication in section 1.3.2.) The reason for this is that if we say "p only if q", then we are saying it is impossible for p to be true unless q is also true, so that if q is false, so is p: that is, $\neg q \rightarrow \neg p$, or in other words, $p \rightarrow q$. The best way to remember this is probably to think of "only if" as meaning "implies". Don't let the position of the word "if" confuse you!

There is a lot of useful advice in section 3.3 of the Alberta Notes—you might want to refer to that as well if you have trouble with the next set of exercises.

1.4.1 Translation exercise

Symbolize the following sentences (use appropriate abbreviations for the various statements, such as H for "Harry will run for class president", *etc.*):

1. Harry and Judith will both run for class president.
2. Either Harry will run for class president or Judith won't.
3. If Harry runs for class president, then Judith won't run, but if Harry doesn't run, then Judith will.
4. If Harry and Judith both run, then George won't run.
5. It won't happen that Harry and Judith both run.
6. If Judith runs, then either Harry won't run or George will.
7. Harry will run if and only if Judith runs.
8. Neither Judith nor Harry will run.
9. George will run, and if Judith runs then Harry will also.
10. Harry will run only if Judith doesn't.
11. Harry will run unless Judith runs.
12. On the assumption that Harry will run if George does, it follows that Judith won't run.
13. Supposing that George runs provided that Judith does, it follows that Harry will run if Judith doesn't.

14. Alcohol and marijuana are drugs.
15. Alcohol and Benzedrine are a deadly combination.
16. Though he loved her, he left her.
17. Cigarettes and whiskey and wild, wild women will drive me crazy.
18. If we reduce pollution and population doesn't increase, our standard of living will not decline, but if we fail to reduce pollution, or if the population increases, then our standard of living will decline.
19. If we fail to reduce population or if the population increases, then our standard of living will decline and we'll have only ourselves to blame.

1.4.2 More translation exercises

For each of the following, construct a truth-functional paraphrase, and symbolize it in propositional logic. Use the following abbreviations:

A Albert jogs regularly.
B Bob jogs regularly.
C Carol jogs regularly.
L Bob is lazy.
M Carol is a marathon runner.
H Albert is healthy.

1. If Bob jogs regularly, he is not lazy.
2. If Bob is not lazy, he jogs regularly.
3. Bob jogs regularly if and only if he is not lazy.
4. Carol is a marathon runner only if she jogs regularly.
5. Carol is a marathon runner just in case she jogs regularly.
6. If Carol jogs regularly, then if Bob is not lazy he jogs regularly.
7. If both Carol and Bob jog regularly, then Albert does too.
8. If either Carol or Bob jogs regularly, then Albert does too.
9. If either Carol or Bob does not jog regularly, then Albert doesn't either.
10. If neither Carol nor Bob jogs regularly, then Albert doesn't either.
11. If Albert is healthy and Bob is not lazy then both jog regularly.
12. If Albert is healthy, he jogs regularly just in case Bob does.
13. Assuming Carol is not a marathon runner, she jogs regularly if and only if Albert and Bob both jog regularly.

14. Although Albert is healthy he does not jog regularly, but Carol does jog regularly if Bob does.
15. If Carol is a marathon runner and Bob is not lazy and Albert is healthy, then they all jog regularly.
16. If Albert jogs regularly, then Carol does provided that Bob does.
17. If Albert jogs regularly if Carol does, then Albert is healthy and Carol is a marathon runner.
18. If Albert is healthy if he jogs regularly, then if Bob is lazy he doesn't jog regularly.
19. If Albert jogs regularly if either Carol or Bob does, then Albert is healthy and Bob isn't lazy.

Now the reverse translation process: using the same abbreviations above, construct natural English sentences whose meaning is given by the following sentences of propositional logic.

1. $A \vee (B \vee C) \to A \wedge (B \wedge C)$
2. $C \to (A \wedge \neg B)$
3. $B \leftrightarrow (\neg L \wedge A)$
4. $\neg A \to (\neg B \to \neg C)$
5. $\neg A \wedge (B \leftrightarrow \neg L)$

1.5 Knights and Knaves

And now for something completely different . . . [24]

A story: There is an island far off in the Pacific, called the island of Knights and Knaves. On this island, there are people called **knights** (who always tell the truth, meaning everything that a knight says must be true) and **knaves** (who always lie, meaning everything a knave says must be false). They may be either male or female. The people of this island are often called "**knavghts**".[25] So some knavghts are knights, some are knaves, and every knight or knave is a knavght.

Another peculiarity of the knavghts: they seem to speak English, but with a small difference, in that they use the connectives of propositional logic in the strict sense we have defined earlier. So they only use "or" in the inclusive sense, and they only use "if . . . then . . . " (and similar expressions) to mean material implication. There is no ambiguity in their use of these propositional connective words.

One day you visit the island: you are then the only non-knavght on the island (so everyone else is either a knight or knave). Suppose you meet two knavghts, and one says, pointing to the other, "He said he was a knave". What can you conclude from this?

Well, clearly the speaker is a knave. You can figure this out this way: no knight could say "I am a knave", for that would be false, and knights always tell the truth. But no knave could say "I am a knave" either, for such a statement would then be true, but a knave always lies. So, no

[24] The material in this section comes (sometimes slightly modified) from the wonderful logic puzzle books of Raymond Smullyan, in particular his book *What is the name of this book?*. This book is, I believe, presently out-of-print, but if you ever come across a copy, I really recommend you buy it—it's a great collection of amusing logic puzzles, of which the following is merely a sample.

[25] "Knavght" is usually pronounced "knot", although folks from Brooklyn sometimes pronounce it "knate".

knavght could ever say "I am a knave", and anyone who tells you otherwise must be telling you a lie. So the speaker told a lie (made a false statement): he must be a knave therefore.

You can see from this analysis that it's often possible to deduce facts about knavghts from statements they make, facts that aren't explicitly part of their statements. In the situation above, the knavght made a statement about his companion, but really he was telling you something about himself (we still don't know whether the companion was a knight or knave).

Let's consider another situation: A knavght man was asked (about his wife, who was also a knavght, and himself) which, if either, was a knight and which, if either, was a knave. He answered "We are both knaves"; what are they?

See if you can figure out the answer yourself before you read the next paragraph!

He cannot be a knight, since a knight couldn't say he was a knave. So he is a knave. Now you might wonder about that, since a knave also couldn't say he's a knave. But that's not really what he said: he said "We are both knaves", which can in fact be a legitimately false statement (such as all knaves always make), provided his wife is not a knave. Careful now: if his wife were a knave, then "We are both knaves" would be true, and so an impossible utterance by a knave. So the only possibility is that he is a knave, his statement is false, and so his wife is a knight.

Here are some other situations; see if you can answer the questions posed. I have given you the answers, with some hints as to how they may be obtained. But try them yourself first.

1. Another knavght man was asked, of his wife and himself, "Are you both knaves?". He answered "At least one of us is"; what are they?

   (Ans: He cannot be a knave, because if he were a knave, his statement would be true, which is impossible for a knave. So he's a knight, and so his statement is true, so his wife must be a knave.)

2. Same situation: this time the man answers "If I am a knight, then so is my wife". What are they?

   (Ans: Assume he's a knight. Then it would follow that his wife is a knight too, since that's just what he said, and if he's a knight, his statement must be true. But look what we have here: we just showed that if he's a knight, then so is his wife. This is exactly what he claimed, and we've just seen this statement is true. Since he said a true statement, he must be a knight, and so therefore his wife must be too. There is a general principle at work here, which I'll summarize below, but see if you can guess what it must be.)

3. Same situation: this time his answer is "My wife and I are of the same type" (meaning either both knights or both knaves). What are they?

   (Ans: You cannot determine the husband's type, he could be knight or knave, but since we know he cannot claim he's a knave, his wife couldn't be a knave since that would in effect mean his statement would be claiming he's a knave too—so she's a knight. You can verify this by cases if you like. Again, there's a general principle working here too.)

There are some general principles which one can see in looking at these situations and their analyses.

1. No knavght can say "I am a knave"; every knavght must claim "I am a knight".

2. For any statement P, if a knavght says "If I am a knight, then P", then the knavght is in fact a knight, and P is true.

3. If a knavght says "If P then I am a knave", then P must be false and the knavght is in fact a knight. (Exercise: this is essentially the same as the previous principle.)

4. If a knavght says "I am a knight if and only if P", then P must be true (but the knavght could be either knight or knave).

5. If a knavght is asked "Is the statement you are a knight equivalent to the statement P?", then a "yes" answer means P is true, and a "no" answer means P is false.

6. Remember that sometimes it's simpler to transform the sentence by standard equivalences, such as $A \rightarrow B \leftrightarrow \neg A \vee B$, $\neg(A \rightarrow B) \leftrightarrow A \wedge \neg B$, $\neg(A \vee B) \leftrightarrow \neg A \wedge \neg B$, $\neg(A \wedge B) \leftrightarrow \neg A \vee \neg B$.

   So, for example: instead of thinking "If P then I am a knave", think "either not P or I am a knave". (Whatever makes it easier for you to deconstruct the sentence.)

You may use these principles in analysing other scenarios, in particular, in solving the following exercises. (If you have trouble, you might like to look at the next section, 1.5.2.)

1.5.1 Knights and knaves exercises

1. We have three people A, B, and C on the Island of Knights and Knaves. Suppose A and B say the following:

   A: All of us are knaves.
   B: Exactly one of us is a knave.

   Can it be determined what B is? Can it be determined what C is?

2. Suppose A says, "I am a knave but B isn't." What are A and B?

3. We again have three inhabitants, A, B and C, each of whom is a knight or a knave. Two people are said to be of the same type if they are both knights or both knaves. A and B make the following statements:

   A: B is a knave.
   B: A and C are of the same type.

   What is C?

4. Again three people A, B and C. A says "B and C are of the same type." Someone then asks C, "Are A and B of the same type?" What does C answer?

5. We have two people A, B, each of whom is either a knight or a knave. Suppose A makes the following statement: "If I am a knight, then so is B." Can it be determined what A and B are?

6. Someone asks A, "Are you a knight?" He replies, "If I'm a knight, then I'll eat my hat!" Prove that A has to eat his hat.

7. A says, "If I'm a knight, then two plus two equals four." Is A a knight or a knave?

8. A says, "If I'm a knight, then two plus two equals five." What would you conclude?

9. Given two people, A, B, both of whom are knights or knaves. A says, "If B is a knight then I am a knave." What are A and B?

10. Two individuals, X and Y, were being tried for participation in a robbery. A and B were court witnesses, and each of A, B is either a knight or a knave. The witnesses make the following statement:

    A: If X is guilty, so is Y.
    B: Either X is innocent or Y is guilty.

    Are A and B necessarily of the same type? (i.e. either both knights or both knaves.)

11. On the island of knights and knives, three inhabitants A,B,C are being interviewed. A and B make the following statements:

    A: B is a knight.
    B: If A is a knight so is C.

    Can it be determined what any of A, B, C are?

12. Another three inhabitants, A, B, C, make these statements:

    A: B is a knave.
    B: A is a knave.
    C: Both A and B are knaves.

    Can it be determined what any of A, B, C are?

13. Suppose the following two statements are true: (1) I love Betty or I love Jane. (2) If I love Betty then I love Jane. Does it necessarily follow that I love Betty? Does it necessarily follow that I love Jane?

14. Suppose that I am a knight, and someone asks me, "Is it really true that if you love Betty then you also love Jane?" I reply, "If it is true, then I love Betty." Does it follow that I love Betty? Does it follow that I love Jane?

15. This problem, though simple, is a bit surprising. Suppose it is given that I am either a knight or a knave. I make the following two statements:

    (a) I love Linda.
    (b) If I love Linda then I love Kathy.

    Am I a knight or a knave?

16. Is There Gold on This Island? On a certain island of knights and knaves, it is rumored that there is gold buried on the island. You arrive on the island and ask one of the natives, A, whether there is gold on this island. He makes the following response: "There is gold on this island if and only if I am a knight." Our problem has two parts:

    (a) Can it be determined whether A is a knight or a knave?
    (b) Can it be determined whether there is gold on the island?

17. Suppose, instead of A having volunteered this information, you had asked A, "Is the statement that you are a knight equivalent to the statement that there is gold on this island?" Had he answered "Yes," the problem would have reduced to the preceding one. Suppose he had answered "No." Could you then tell whether or not there is gold on the island?

18. The First Island. On the first Island he tried, he met two natives A, B, who made the following statements:

    A: B is a knight and this is the island of Maya.
    B: A is a knave and this is the island of Maya.

    Is this the island of Maya?

19. The Second Island. On this Island, two natives A, B, make the following statements:

    A: We are both knaves, and this is the island of Maya.
    B: That is true.

    Is this the island of Maya?

20. The Third Island. On this island, A and B said the following: A: At least one of us is a knave, and this is the island of Maya. B: That is true. Is this the island of Maya?

21. Here is a bit of an offbeat question. One day, on the island of Knights and Knaves, you see an inhabitant. You go up to her and ask: "Are you a knight or are you a knave?" She says: "I won't tell you" and walks away. Is it possible to decide if she is a knight or a knave?

1.5.2 What's it all about?—more general principles

There are several serious points about the knights and knaves story (sorry! it isn't all fun and games after all!), which have to do with how negation acts with the various connectives. We shall see this in several contexts, but here are a few comments to go on.

Notice there is a difference between a knavght saying one sentence: "p and q." and a knavght saying two sentences: "p." "q." We saw that early on: consider the difference between a knave saying "We are both knaves." (referring to himself and his wife), and a knave saying "I am a knave. My wife is a knave." The second utterance is impossible, since he cannot say "I am a knave" (it would be a true statement uttered by a knave, an impossibility). But the first statement is possible: "We are both knaves." (which is the same as "I am a knave and my wife is a knave.") could be said by a knave, provided his wife is a knight. This reflects the fact that the negation of a conjunction is a disjunction: $\neg(p \wedge q) \leftrightarrow \neg p \vee \neg q$. In the language of knights and knaves, a knave saying "$p \wedge q$" means at least one of p, q is false (maybe both, maybe not). It does not mean both p and q *must* be false. But that's just what a knave saying "p. q." amounts to: both p and q would then have to be false. This distinction would not hold for knights, for they always tell the truth, and "$p \wedge q$" is true precisely if both p and q are true.

We can summarize this sort of thing as follows.

If a knight says	then	if a knave says	then
$\neg p$	p is $\bot$	$\neg p$	p is $\top$
$p \wedge q$	both p and q are $\top$	$p \wedge q$	at least one of p or q is $\bot$
$p \vee q$	at least one of p or q is $\top$	$p \vee q$	both p and q are $\bot$
$p \longrightarrow q$	either p is $\bot$ or q is $\top$	$p \longrightarrow q$	p must be $\top$ and q must be $\bot$

Be sure you understand this—it will help in doing the exercises of course, but it also should help firm up your understanding of how the connectives work.

[Optional:] Translation into propositional logic

Finally, it is actually possible to translate a knight/knave problem into pure propositional logic (I don't really suggest you do this to solve knights and knaves problems, but it is an interesting observation).

The crucial point is that A can make a statement P if and only if the statement " 'A is a knight' is equivalent to P" is true.

To see why this is so, consider first what it means for A to assert P: if A is a knight, P must be true, and if P is true, then (since A said P, *i.e.* the truth) A must be a knight. On the other hand, if " 'A is a knight' is equivalent to P" is true, then if A is in fact a knight, P must be true, and so A can say P, whereas if A is in fact a knave, then P must be false, and so again, A can say P.

Let's abbreviate "A is a knight" by simply A, and "A is a knave" by $\neg A$ ("A is not a knight"), so that the statement "A says P" is equivalent to the statement $A \leftrightarrow P$. This is notationally dubious, since we are using the same letter A to mean two totally different things: a person (the knavght making the assertion), and a statement (that he is a knight). It is convenient, however, for we have just shown that with this abbreviation, we can read $A \leftrightarrow P$ as "A says P" as well as "A is equivalent to P", making these propositional formulas a bit easier to read.

This allows us to translate statements about the statements of knavghts into statements in propositional logic. For example, consider our second example, where a knavght man was asked about his knavght wife and himself which, if either, is a knight and which, if either, is a knave, and he answered "We are both knaves". Translating this, we would get the following: $A \leftrightarrow (\neg A \wedge \neg B)$, where A is the man and B is his wife. The solution we outlined essentially amounted to showing this implies $\neg A \wedge B$. In other words, we have to show $[A \leftrightarrow (\neg A \wedge \neg B)] \longrightarrow [\neg A \wedge B]$ is a tautology, which is a standard exercise in truth tables. (In fact, the $\longrightarrow$ can be strengthened to $\leftrightarrow$, and we'd still have a tautology. In other words, if A is a knave and his wife is a knight, then if asked what they are, A could[26] reply "We are both knaves". Check this by analysing the possibilities.)

Optional exercise: Verify all this, and translate some of the other examples and statements of principle.

By the way: the usual classic knights-and-knaves puzzle is this: going to the town of EverlastingDelights, you come to a fork in the road, where you meet a knavght. Not sure which way to go, you want to ask him which direction will get you there; what single question, with only "yes" or "no" as possible answers, could you put to him which will allow you to know what direction to go?

There are some variants of this: here are two.

Two knavghts are standing at a fork in the road. By asking one yes/no question to one of them, can you determine the direction to the town of EverlastingDelights? And, by asking one yes/no question to one of the knavghts, can you determine whether he is a knight?

[26] There are other replies he could also make—find as many as you can.

1.6 Answers to the exercises

Exercises 1.2.4

There is considerable room for variant answers—if you have questions about these answers, ask them!

BTW: these are clearly not arguments in propositional logic (or rather, to make them so, many more premises would have to be added, premises having to do with attitudes and so on).

In each case, I have put the premises (numbered) above the horizontal line, and the conclusion below it.

1.

1. Perfection of soul corrects the inferiority of the body
2. Physical strength without intelligence does nothing to improve the mind

---

It is right that men should value the soul rather than the body

2.

1. What is empty is nothing
2. What is nothing cannot be

There cannot be any emptiness

3.

1. The subject of the gods' existence and form is obscure
2. Human life is short

About the gods, I am not able to know whether they exist or do not exist, nor what they are like in form

4.

1. Other creatures are soon self-supporting
2. Man alone needs prolonged nursing

In the beginning man was born from creatures of a different kind

5.

1. Either death is a state of nothingness and utter unconsciousness, or there is a change and migration of the soul from this world to another

---

Death is good

Exercise 1.3.6

1: (a) negation (b) conjunction (c) implication (d) disjunction

(I'll leave 2, 3 to you.)

Exercise 1.3.8

(a) everything (b) everything (c) 7, 8, 9, 10 (d) 2, 3, 4, 5, 6 (e) 5, 6 (f) 7, 8, 9, 10
(g) 8, 9 (h) 3, 4 (i) none (j) 4 (k) 6 (l) 5, 6 (m) 9 (n) 7, 10 (o) 8, 9

Exercise 1.3.10

It will be clear from number 3 that the main column for numbers 1, 2 will be the same as the main column for the biconditional ($\top$, $\bot$, $\bot$, $\top$). Use this to check your answers.

Here's number 4: (The column with the compound formula's truth values is indicated by a $*$.)

p	q	$\neg p$	$\wedge$	q
$\top$	$\top$	$\bot$	$\bot$	
$\top$	$\bot$	$\bot$	$\bot$	
$\bot$	$\top$	$\top$	$\top$	
$\bot$	$\bot$	$\top$	$\bot$	
			*	

p	q	$\neg$	$(p \wedge q)$
$\top$	$\top$	$\bot$	$\top$
$\top$	$\bot$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\top$	$\bot$
		*	

p	q	$\neg p$	$\wedge$	$\neg q$
$\top$	$\top$	$\bot$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\bot$	$\top$
$\bot$	$\top$	$\top$	$\bot$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

p	q	$\neg p$	$\vee$	$\neg q$
$\top$	$\top$	$\bot$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\top$
$\bot$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

Clearly the only two that are equivalent are $\neg(p \wedge q)$ and $\neg p \vee \neg q$.

Exercise 1.3.15

1.

A	B	$(A \rightarrow B)$	$\rightarrow$	$\neg A$
$\top$	$\top$	$\top$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

2.

A	B	$(A \wedge B)$	$\wedge$	$\neg A$
$\top$	$\top$	$\top$	$\bot$	$\bot$
$\top$	$\bot$	$\bot$	$\bot$	$\bot$
$\bot$	$\top$	$\bot$	$\bot$	$\top$
$\bot$	$\bot$	$\bot$	$\bot$	$\top$
			*	

3.

A	B	A	$\wedge$	$\neg$	$(B \vee A)$
$\top$	$\top$	$\top$	$\bot$	$\bot$	$\top$
$\top$	$\bot$	$\top$	$\bot$	$\bot$	$\top$
$\bot$	$\top$	$\bot$	$\bot$	$\bot$	$\top$
$\bot$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$
			*		

4.

A	B	$(A \rightarrow B)$	$\leftrightarrow$	$(\neg B$	$\rightarrow$	$\neg A)$
$\top$	$\top$	$\top$	$\top$	$\bot$	$\top$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\top$	$\bot$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\bot$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$
			*			

5.

A	B	$(A \rightarrow B)$	$\rightarrow$	$(B$	$\vee$	$\neg A)$
$\top$	$\top$	$\top$	$\top$	$\top$	$\top$	$\bot$
$\top$	$\bot$	$\bot$	$\top$	$\bot$	$\bot$	$\bot$
$\bot$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\bot$	$\top$	$\top$	$\bot$	$\top$	$\top$
			*			

6.

A	B	$(A \rightarrow B)$	$\vee$	$(B \rightarrow A)$
$\top$	$\top$	$\top$	$\top$	$\top$
$\top$	$\bot$	$\bot$	$\top$	$\top$
$\bot$	$\top$	$\top$	$\top$	$\bot$
$\bot$	$\bot$	$\top$	$\top$	$\top$
			*	

And here is the rest:
(1) "If A implies B then A is false" (contingency)
(2) "A, B and $\neg A$ are all true" (contradiction)
(3) "A is true, but not 'B or A'" (contradiction)
(4) "'A implies B' is equivalent to 'not B implies not A'" (tautology)
(5) "If A implies B then either B is true or A is false" (tautology)
NOTE: It is worth noticing that in fact "A implies B" is actually *equivalent* to "either B or not A".
(6) "Either A implies B or B implies A" (tautology)

Exercise 1.4.1

Variants are possible—check with me if you're not sure.

(1) $H \wedge J$ (2) $H \vee \neg J$ (3) $(H \rightarrow \neg J) \wedge (\neg H \rightarrow J)$ (4) $(H \wedge J) \rightarrow \neg G$ (5) $\neg(H \wedge J)$
(6) $J \rightarrow (\neg H \vee G)$ (7) $H \leftrightarrow J$ (8) $\neg H \wedge \neg J$ (9) $G \wedge (J \rightarrow H)$ (10) $H \rightarrow \neg J$
(11) $H \vee J$ (12) $(G \rightarrow H) \rightarrow \neg J$ (13) $(J \rightarrow G) \rightarrow (\neg J \rightarrow H)$ (14) $A \wedge M$ (15) C
(16) Loved $\wedge$ Left (17) C (18) $(R \wedge \neg I \rightarrow \neg D) \wedge (\neg R \vee I \rightarrow D)$
(19) $(\neg R \vee I) \rightarrow (D \wedge B)$

Exercises 1.4.2

There are possible variants, but I've generally given the one "closest" to the English.

1. $B \rightarrow \neg L$
2. $\neg L \rightarrow B$
3. $B \leftrightarrow \neg L$
4. $M \rightarrow C$
5. $C \leftrightarrow M$ (though possibly $M \rightarrow C$ or $C \rightarrow M$)[27]
6. $C \rightarrow (\neg L \rightarrow B)$
7. $(C \wedge B) \rightarrow A$
8. $(C \vee B) \rightarrow A$
9. $(\neg C \vee \neg B) \rightarrow \neg A$
10. $(\neg C \wedge \neg B) \rightarrow \neg A$
11. $(H \wedge \neg L) \rightarrow (A \wedge B)$
12. $H \rightarrow (A \leftrightarrow B)$ (though possibly $H \rightarrow (A \rightarrow B)$ or $H \rightarrow (B \rightarrow A)$)[24]
13. $\neg M \rightarrow (C \leftrightarrow (A \wedge B))$
14. $H \wedge \neg A \wedge (B \rightarrow C)$
15. $(M \wedge \neg L \wedge H) \rightarrow (C \wedge B \wedge A)$
16. $A \rightarrow (B \rightarrow C)$
17. $(C \rightarrow A) \rightarrow (H \wedge M)$
18. $(A \rightarrow H) \rightarrow (L \rightarrow \neg B)$
19. $(C \vee B \rightarrow A) \rightarrow (H \wedge \neg L)$

Translations from propositional logic (there are lots of correct variations as well).

1. If any of Albert, Bob, or Carol jog regularly, then they all do.

[27]There is a possible dispute about the meaning of "p just in case q"; on reflection, I lean to it meaning $p \leftrightarrow q$, though an argument could be made to support $p \rightarrow q$ or even $q \rightarrow p$.

2. If Carol jogs regularly, then Albert does but Bob doesn't.
3. Bob jogs regularly if and only if he's not lazy and Albert jogs regularly.
4. If Albert doesn't jog regularly, then Carol doesn't if Bob doesn't.
   (*This is equivalent to:* Carol doesn't jog regularly if neither Albert nor Bob does.)
5. Albert doesn't jog regularly, and Bob jogs regularly if and only if he's not lazy.

Exercise 1.5.1

On a test, you would have to provide your reasons for the answers; here however I have usually merely given the conclusion, with a hint in a few cases. To abbreviate things a bit, I have adopted the following notation: if A is a knavght: $\top(A)$ means "A is a knight"; its negation, $\neg\top(A)$, also denoted $\bot(A)$, means "A is a knave". I hope it's clear why I use this notation: $\top(A)$ not only means "A is a knight", it also means "everything A says is $\top$", and similarly for $\bot$. $?(A)$ means "we do not know the type of A".

1. $\bot(A), ?(B), \top(C)$
2. $\bot(A), \bot(B)$
3. $\bot(C)$ (but $?(A), ?(B)$)
4. "yes"
5. $\top(A), \top(B)$ (Use principle 2)
6. $\top(A)$, so he must eat his hat (Use principle 2)
7. $\top(A)$ (Use principle 2)
8. This statement cannot be made by any knavght—so I must be a knave(!).
9. $\top(A), \bot(B)$ *via* principle 3
10. same type
11. all $\top$
12. $\bot(C)$; A and B are not both the same type (so either $\top(A), \bot(B)$ or $\bot(A), \top(B)$).
13. I love Jane (but ?Betty)
14. I love Betty (but ?Jane)
15. $\top(me)$
16. There is gold on the island (but $?(A)$)
17. No gold on the island (*via* principle 5)
18. $\bot(B), \bot(A)$ so not Maya
19. $\bot(A), \bot(B)$ so not Maya
20. $\bot(A), \bot(B)$ so not Maya
21. She's a knight.

1.7 X-treme Knights and Knaves—getting a bit blood-thirsty!

A visit to Transylvania

More problems from Smullyan's *What is the name of this book?*

Preliminaries

In Transylvania, the population is made up of humans (who always intend to tell the truth) and vampires (who always intend to lie); what complicates the matter is that half the population is insane: they believe every true statement is really false, and *vice versa*. So, sane humans and insane vampires always tell the truth, but insane humans and sane vampires always lie. (For example, an insane vampire intends to lie, but since he thinks true statements are false and false statements are really true, he ends up actually telling the truth.)

So, let's explore some of the consequences of this odd situation. For example, if a Transylvanian says "I am not a sane human", what can you conclude? He cannot actually be a sane human (for if so, he'd say so), nor can he be an insane human (for, being a human, he'd want to tell the truth, but couldn't, so he'd say he was not an insane human, or equivalently, he was either sane or a vampire—he wouldn't say he was not a sane human, for that would be equivalent to saying he was either insane or a vampire, which is a true statement in his case). Check for yourself he cannot be a sane vampire either, so he must be an insane vampire.

Another example (showing an alternate way to look at such statements): suppose he said "I am human, or I am sane". If this statement is false, he must be an insane vampire, so his statement must be true, which is a contradiction. So the statement is true, and so he's either human or sane, but also he must be (because he told the truth) either a sane human, or an insane vampire. Only a sane human fits both conditions. So he must be a sane human.

Here are some to try for yourself (the answers are in the footnotes). If he said "I am an insane human", what is he?[28] If he said "I am a vampire", what can you conclude?[29] If he said "I am insane", what can you conclude?[30]

Here's an interesting principle: If a Transylvanian believes that he believes something, then that something must be true. If he does not believe that he believes something, then that something must be false. (Note that his merely believing something doesn't tell you about its truth or falsehood—it's the believing that he believes it that is crucial here!) Try to convince yourself of this principle.

Here is an even more important principle: If a Transylvanian says "I believe X", where X is some statement, then if he is human, X must be true, whereas if he is a vampire, then X must be false. (Convince yourself of this!)

Problem 1: I meet two Transylvanians, A and B. I ask A "Is B human?", and A replies "I believe so." I ask B "Do you believe A is human?" B answered yes or no; which was it: "yes" or "no"?[31]

Another principle, an old one this time: Let's call sane humans and insane vampires "knightlike", and insane humans and sane vampires "knavelike" (for the obvious reasons). Then, if a knightlike individual says "If I am knightlike, then X" (for some statement X), then he must be knightlike, and X must be true.

[28] He's a sane vampire.
[29] He's insane (but could be human or vampire).
[30] He's a vampire (but could be sane or insane).
[31] "Yes"

By the way: if you asked a Transylvanian “Are you knightlike?” what would his answer be?[32] If you asked a Transylvanian “Do you believe you are knightlike?” what can you conclude from his answer?[33]

Is Dracula alive and well in Transylvania?

Any tourist to Transylvania is bound to ask himself this question; suppose you asked a Transylvanian about this, and he replied “If I am human, then Count Dracula is still alive”, then what can you conclude? ... Well, think it over: you should realize that you still won’t know what you want, even if you asked a knightlike Transylvanian (he could be a sane human, so Dracula would be alive, or he could be an insane vampire and Dracula might be alive—or dead!). Check that the same indeterminacy holds if you get the answer “If I am sane, then Count Dracula is still alive”, or even if you get the answer “If I am a sane human, then Count Dracula is still alive”. However, if you get the answer “If I am either a sane human or an insane vampire, then Count Dracula is still alive”, then you will definitely know Dracula is really alive (because then he is saying he is knightlike, which only knightlike individuals can do—we saw a similar principle when we were doing ordinary knights and knaves problems).

Can you think of a statement you might receive as an answer that would convince you that (a) Dracula is alive and (b) the statement itself is false. How about an answer-statement which would convince you that Dracula is alive but for which you couldn’t determine if the statement is true or false?[34]

Problem 2: Suppose a Transylvanian made these statements:

(1) I am sane.
(2) I believe that Count Dracula is dead.

Can you determine whether Dracula is alive?

Problem 3: Suppose instead that the Transylvanian made these statements:

(1) I am human.
(2) If I am human then Count Dracula is alive.

Can you determine whether Dracula is alive?

Problem 4: Here are some quickies: Find a single question you can ask a Transylvanian which will determine whether he is a vampire or not. Now find one to determine if he is sane or not. Next, find one which will force him to answer “Yes”, regardless of what sort of individual he is. And finally, find a question which will determine if Count Dracula is still alive.[35]

Dracula’s Castle

Now things get interesting: the upper aristocracy in Transylvania use the old traditional language for some words, and in particular, they don’t use “yes” and “no”, but instead “bal” and “da”—the problem is, you don’t know which means which! So, when one day you find yourself

[32] “Yes”, regardless of what type of Transylvanian he is.

[33] “Yes” would mean he was sane; “no” would mean he was insane.

[34] “I am knavelike and Dracula is dead”; “I am knightlike if and only if Dracula is still alive”. There are other possible answers. Show that another answer for the second situation is “I believe that if someone asked me whether Dracula was alive, I would answer ‘Yes’ ”.

[35] **2.** Dracula is dead. **3.** Dracula is alive. **4.** “Are you sane?”; “Are you human?”; “Do you believe you are human?” or “Are you knightlike?”; “Is the statement that you are knightlike equivalent to the statement that Dracula is alive?” or “Do you believe that the statement that you are human is equivalent to the statement that Dracula is alive?”

invited to Dracula's Castle, which is inhabited by aristocrats (possibly including Dracula himself!), you have to deal with the situation that not only do you not know which type of individual everyone is, but you cannot really tell what they're answering when they say "bal" or "da". But think about it a bit: it is possible to ask a single question (which will get you a "bal/da" answer) which will tell you whether the individual you are speaking to is a vampire: namely "Is 'Bal' the correct answer to the question 'Are you sane?'?" (Check this for yourself!)

Now find similar questions which will determine whether you are speaking to an individual who is sane or not; to determine what "Bal" means; to force him to answer "Bal" to your question; and to find out whether or not Dracula is alive.[36]

Dracula's Challenge

Finally, you get to meet Dracula. Normally, this wouldn't be a *good thing*, but he offers you a chance to save your life: he points out that although you've been very clever to get all those questions to sort out who's who, you've missed a general underlying principle that solves all such questions. If you can find that principle, he will let you go away unharmed (but if you cannot, then he'll turn you into a vampire!). There is one single sentence S with the miraculous property that it can help you determine the truth of any other sentence X merely by asking any individual "Is S equivalent to X?". For if they answer "Bal", X must be true, but if they answer "Da", then X must be false. (For example, to find out if Dracula is alive, you'd just have to ask any Transylvanian aristocrat "Is S true if and only if Dracula is alive?") To save your life, you must tell me what the sentence S is. (!)

Find a good candidate for S, to save your life.[37]

[36] "Is 'Bal' the correct answer to 'Are you human?'?" (If he answers "Bal" then he's sane.) "Do you believe you are human?" (Everybody must answer "Yes" to that, so whatever he says must mean "yes".) "Is 'Bal' the correct answer to the question 'Are you knightlike?'?" (Another question would be "Are you knightlike if and only if 'Bal' means 'Yes'?" Both these will force an answer "Bal".) And finally, "Do you believe that 'Bal' is the correct answer to the question 'Is the statement that you are human equivalent to the statement that Dracula is alive?'?", or similarly "Is 'Bal' the correct answer to the question 'Is the statement that you are knightlike equivalent to the statement that Dracula is alive?'?"

[37] Call a Transylvanian "Bal-ish" if he answers 'Bal' to the question 'Is $1+1=2$?'?" (or any similarly always-true question). The point is: if "Bal" means "Yes", then Bal-ish Transylvanians are knightlike, and if "Da" means "Yes", then Bal-ish Transylvanians are knavelike. So, $S =$ "You are Bal-ish" does the trick: If the answer to "Is X equivalent to the statement that you are Bal-ish?" is "Bal", then X must indeed be true, and X must be false if the answer is "Da". Check this yourself!